Computer Vision and Image Understanding 194 (2020) 102881

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Graph convolutional neural network for multi-scale feature learning ☆

Michael Edwards [a], Xianghua Xie [a,*], Robert I. Palmer [a], Gary K.L. Tam [a], Rob Alcock [b], Carl Roobottom [c]

[a] Department of Computer Science, Swansea University, Singleton Park, Swansea SA2 8PP, United Kingdom
[b] Peninsula Radiology Academy, Plymouth Hospitals NHS Trust, Plymouth PL6 5WR, United Kingdom
[c] Plymouth University Schools of Medicine & Dentistry, Plymouth Hospitals NHS Trust, Plymouth PL6 8BT, United Kingdom

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Deep learning
Graph convolutional neural network
Medical image segmentation
Marginal space learning
Aortic root
Computerized tomography

ABSTRACT

Automatic deformable 3D modeling is computationally expensive, especially when considering complex position, orientation and scale variations. We present a volume segmentation framework to utilize local and global regularizations in a data-driven approach. We introduce automated correspondence search to avoid manually labeling landmarks and improve scalability. We propose a novel marginal space learning technique, utilizing multi-resolution pooling to obtain local and contextual features without training numerous detectors or excessively dense patches. Unlike conventional convolutional neural network operators, graph-based operators allow spatially related features to be learned on the irregular domain of the multi-resolution space, and a graph-based convolutional neural network is proposed to learn representations for position and orientation classification. The graph-CNN classifiers are used within a marginal space learning framework to provide efficient and accurate shape pose parameter hypothesis prediction. During segmentation, a global constraint is initially non-iteratively applied, with local and geometric constraints applied iteratively for refinement. Comparison is provided against both classical deformable models and state-of-the-art techniques in the complex problem domain of segmenting aortic root structure from computerized tomography scans. The proposed method shows improvement in both pose parameter estimation and segmentation performance.

1. Introduction

The use of contextual and local information is common in numerous domains, from understanding scene context in images to modeling sentence structure within speech (Yu et al., 2016; Oh et al., 2017). The idea of a scale-space is introduced by Koenderink (1984) and discussed by Lindeberg (1996) and Tony (2008), in which a multi-resolution decomposition of an input signal is an ordered set of signals at increasingly coarser representations, reducing the finer scale features of an input domain and providing an increasingly generalized representation of the data (Florack et al., 1996). Given that an observed dataset may describe a sampling of a problem domain in which the spatial scale of the target may be unknown, it can be beneficial to represent the observation across multiple scales. Lindeberg (1996) discusses that the use of feature descriptors are often dependent on the relationship between the size of points of interest within the data and the size of the operators which are to be applied to them. Developing feature extractors that are able to provide information from various levels of scale has been an important area of research in vision communities and is closely linked to drop off in resolution for biological vision systems (Lindsay and Norman, 1977; Curcio et al., 1990), as shown by the topology of the photoreceptors within the human eye, Fig. 1. The density of sensory structures within the eye provides a region of focal acuity, whilst the reduction in density towards the outer field of view leads to a reduced resolution. Rosenholtz (2016) explains that the loss of acuity in peripheral vision should not result in the perception of a blurred scene, as appears within Fig. 2, despite the drop in resolution as angular distance increases from the fovea and center of focus. The understanding of scale and the utilization of contextual information is an important task in computer vision and its use in pattern recognition, and methodologies have been explored which look to handle changes in scale and the relation between an object and its wider context.

Current deep learning approaches incorporate scale information by either producing filters of different sizes or input streams of different resolutions. Both approaches keep the local and contextual information separate until they are fused downstream. In order to explore the use of spatial representation learning on incorporated contextual and local features, we present a generalized methodology for constructing a multi-resolution graph using an irregularly spaced patch sampling

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.cviu.2019.102881.

* Corresponding author.
E-mail address: X.Xie@Swansea.ac.uk (X. Xie).

https://doi.org/10.1016/j.cviu.2019.102881
Received 16 October 2018; Received in revised form 2 September 2019; Accepted 26 November 2019
Available online 2 December 2019

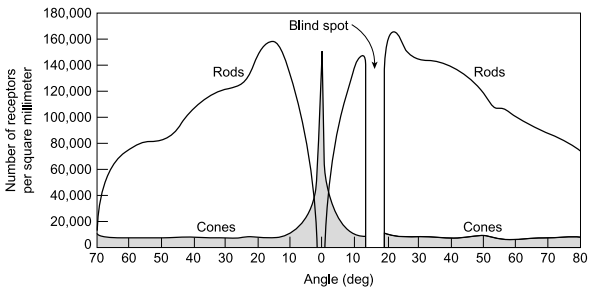


Fig. 1. Spatial resolution of visual sensory receptors within the human eye. The resolution of peripheral vision gradually reduces as the angle of observation deviates from the foveal region (0 degrees).
Source: Image from Lindsay and Norman (1977).

Fig. 2. Effect of drop in acuity within the peripheral visual field. Left to right: Original image, increasing reduction in spatial acuity, leading to a loss of high resolutions within the peripheral sampling. "Blurring" is exaggerated in order to adequately display the effect.
Source: Image from Rosenholtz (2016).

method. By using a novel multi-resolution pooling method to create a relatively small patch which contains both local and contextual structural information, we are able to learn features from raw intensities; avoiding the inefficiencies of large patches and the need to train numerous Convolutional Neural Network (CNN) models for each scale. This Graph-CNN network acts as a detection classifier for a search space optimization framework, which not only eliminates the burden of defining hand-crafted local and contextual features during training, but also significantly reduces the number of potential object pose parameter hypotheses at the testing stage.

The presented study is structured as follows. Section 2 provides a background to the proposed methods. Section 3 introduces an application domain of medical segmentation, on which we evaluate the multi-resolution sampling and deep learning on the irregular spatial domain of the non-uniformly sampled grid. Section 4 outlines the proposed pipeline for automated segmentation using deep learning on the irregular domain. Section 5 presents multi-resolution deep feature learning to drive MSL for position-orientation pose parameter estimation, and deformable model segmentation is utilized to obtain accurate regularized meshes. In Section 6 we evaluate the proposed pipeline on the case study of aortic root detection and segmentation on Computerized Tomography (CT) scans of the human torso, providing qualitative and quantitative analysis in Section 7. Conclusions and discussions on the results found are given in Section 8.

2. Related methodology

2.1. Feature scale and contextual information

Many methods, such as Scale Invariant Feature Transform (SIFT) and scale cascades have been developed to utilize hand-crafted descriptor sets in previous years; including the Haar, tilted-Haar, and steerable Haar features seen in many image based object detectors (Lowe, 1999; Bay et al., 2008; Freeman and Adelson, 1991; Viola and Jones, 2001; Lienhart and Maydt, 2002). Cascade based methods, such as those presented in the Viola–Jones cascade detector, Viola and Jones (2001), aim to speed up detection by detecting on contextual features before moving on to local information, utilizing the fast computation of descriptors to discard regions which do not match the learned context.

In more recent years however, the use of deep learning algorithms have become a popular alternative, capable of learning feature descriptors by combining inputs and adjusting their related importance weighting (LeCun et al., 2015). Standard neural networks have been shown to perform well in domains which exhibit no assumption of spatial relation between input features, and recently the usage of CNNs in domains residing on a regular Cartesian grid, such as 2D images and 3D volumes, has shown that spatially localized features can present beneficial descriptors for problems such as object recognition and detection (LeCun et al., 1998; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). Given that the appearance of local structure can significantly vary, contextual structures are often just as important for detection as local details. Using CNNs on large enough 3D patches to capture both local and contextual features is computationally impractical, often requiring complicated networks to capture information at various scales (Cai et al., 2016; Milletari et al., 2016). Kamnitsas et al. (2017) formulated a CNN architecture with multiple branches, one for each resolution, learning spatial features at different resolutions for brain lesion segmentation. Each branch contains its own collection of filters and the learning of high- and low-resolution features are disjoint between the multiple branches. A similar branching scheme is proposed by Kawahara and Hamarneh (2016), with multiple resolutions being kept separate along different tracts of the architecture before being combined as input to a fully-connected architecture. He et al. (2014) proposes a spatial pyramid pooling layer, maintaining local spatial information and removing the need to fix input sizes to a CNN when computing a fixed-length output vector. Ren et al. (2016) combine a object proposal scheme with a CNN classifier to learn spatial information through a region-of-interest pyramid of reference boxes, with a Region Proposal Network identifying key areas for the Fast R-CNN classifier to focus its attention. Fig. 3 gives an overview of current approaches to multi-scale deep learning.

2.2. Marginal space learning

The estimation of pose parameters is often necessary for 3D object detection, for example there may be 3 optimal parameters each for position (x, y, z), orientation (ω, ϕ, θ), and local scale (S_x, S_y, S_z). Detection can often be formulated as a classification problem; however to exhaustively represent or search all pose combinations in a single high-dimensional space, Ψ, is computationally impractical. Most anatomical structures have some natural alignment (i.e. the aortic root is near the left ventricle) and therefore it is observed that the probability distribution is clustered in a small localized region of Ψ. The idea of MSL, Zheng and Comaniciu (2014), is that the full similarity search space can be marginalized in an attempt to reduce complexity for each increasing level of pose estimation:

$$\Psi_a \subset \Psi_{ab} \subset \Psi_{abc} = \Psi\,, \tag{1}$$

where Ψ_a is the position search space, Ψ_{ab} is the position-orientation space, and Ψ_{abc} is the position-orientation-scale space. It is assumed that the optimal hypothesis Π is contained within the highest probability hypotheses of all marginal spaces, such that

$$\Pi = \Pi_{abc} \subset \Pi_{ab} \subset \Pi_a\,. \tag{2}$$

Given three marginal spaces in (1), three classifiers C_a, C_{ab} and C_{abc} can be trained. At the testing stage, C_a can eliminate the vast majority of false hypotheses in Ψ_a, leaving high probability hypotheses Π_a. These are then passed through C_{ab} to leave Π_{ab}, which are subsequently

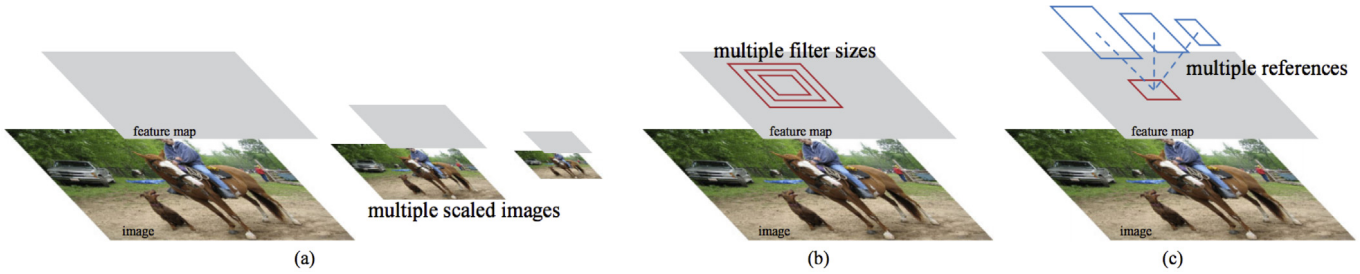


Fig. 3. Current multi-scale schemes in feature representation learning in deep learning approaches. (a) Pyramids of images and filters (branching), (b) Pyramids of filters, (c) Pyramids of reference boxes.
Source: Adapted from Ren et al. (2016).

passed through C_{abc} to leave $\Pi_{abc} = \Pi$. MSL therefore dramatically alleviates the high computation needed for exhaustive search and has been shown to reduce the number of test hypotheses significantly for applications in 3D volumes (Zheng et al., 2012; Palmer et al., 2015). When training both position and orientation estimator models, a positive hypothesis must satisfy the condition that

$$|P_k - P_k^t|/S_k \leq 1 \quad \forall k\,, \tag{3}$$

where P_k is a single pose hypothesis, P_k^t is its ground truth, and S_k is the corresponding parameter search step.

Formulation of the problem is made as a binary classification task over a regression approach in order to obtain probabilities for selection of a set of hypothesis regions, decreasing the search space without confining the search too much. Deep learning approaches to pose estimation from medical volumes have seen a recent advance (Gessert et al., 2018), however MSL is approached as a binary hypothesis detection problem. It is possible to utilize a logistic regression for the binary classification, however the non-linear activations of a network architecture are able to model more complex boundary decisions and as such can often more accurately reflect more complex problem domains when avoiding local minima traps (Goodfellow et al., 2016).

2.3. Graph-based convolutional neural networks

Due to the irregular spatial domain that is provided by a sampling operator which contains information from a non-uniform multi-resolution operation, it is non-trivial to apply standard CNN operators to the patches. We can use fully-connected NN models to learn features for our marginal space classifiers, however such architectures have no constraint on the spatial localization of features from the input space, as would be learned by CNN classifiers on the regular Cartesian domain. Due to their assumption of the regular kernel as a sampling operator, such CNN operators are ill-defined for use on non-euclidean domains. To make use of spatial relationships between the input features in an irregular domain we can formulate the sampling operator as an irregularly spaced sampling of the underlying domain and by defining the multi-resolution patch topology as a graph G, with the input intensities forming a graph signal f that resides on G we can utilize the graph Laplacian as a method of encoding the underlying spatial topology of the domain for the purpose of defining localized information. By utilizing Graph-CNN operators, spatially localized features can be learned via spectral filtering techniques developed in the field of deep learning and signal processing on graphs (Shuman et al., 2013; Henaff et al., 2015; Defferrard et al., 2016; Kipf and Welling, 2017; Edwards and Xie, 2016). This allows end-to-end learning of features on the irregular space that allow our model to simultaneously observe local features and low-resolution wider contextual information without the overhead of learning a different CNN model for each scale. The utilization of deep learning approaches which utilize graph structures is growing rapidly in recent years, including the learning of graph-wise signals that reside on the graph (Bruna et al., 2014; Edwards and Xie, 2016, 2017), node-wise segmentation (Defferrard et al., 2016; Kipf and Welling, 2017; Monti et al., 2016; Qi et al., 2017; Wang et al., 2018) and graph structure learning (Ying et al., 2018)

From graph construction we obtain the edge weighting and diagonalized adjacency matrices, $W \in \mathbb{R}^{N\times N}$ and $A \in \mathbb{R}^{N\times N}$ respectively. This allows us to construct the non-normalized Laplacian matrix representation of the graph structure, $L \in \mathbb{R}^{N\times N}$, by $L = D - W$. Given a complete Laplacian decomposition we can formulate a Fourier basis for the graph domain and utilize the Graph-CNN operators, comprised of the eigenvector matrix $U \in \mathbb{R}^{N\times N}$ ordered by its associated decreasing eigenvalues $\lambda \in \mathbb{R}^N$. Such a basis allows us to represent a given graph signal $f \in \mathbb{R}^N$ in the spectral frequency domain of the graph by computing the spectral signal $\hat{f}$ via the Graph Fourier Transform (GFT):

$$\hat{f} = U^T f\,. \tag{4}$$

By defining a convolution in the spatial domain as the element-wise multiplication in the frequency domain, we are able to produce spectral filtering operations on $\hat{f}$ with a set of spectral multipliers k by

$$f * k = \hat{f} \odot \hat{k} \tag{5}$$

where $*$ defines the convolutional operation, and $\odot$ represents the element-wise product. Optimizing the weightings of spectral multipliers via back-propagation allows the training of a self-learning feature mining architecture, rather than arduously defining hand-crafted features for a complex domain topology. To ensure that localized features are learned in the spatial domain, we can utilize the property of smoothness in the frequency domain providing spatial locality on the spatial domain (Henaff et al., 2015; Edwards and Xie, 2016); thus the network tracks $n < N$ weights for each filter, interpolating them up to N for use in (5).

Another core operation in the CNN architectures is the use of pooling, striding across the regular spatial domain of the input feature map with an appropriate max or mean operator, to produce a coarsened resolution map as output. Such pooling operations provide two main benefits, firstly the memory and computational complexity for convolution is reduced for smaller sized feature maps, secondly the learned features are generalized by compression of feature map resolution (Boureau et al., 2010). The standard CNN pooling operator maintains the spatial regularity of the domain, taking a Cartesian grid as input and returning a Cartesian grid feature map as output. The use of element-wise spectral filtering within the Graph-CNN convolutional operator means each layer of a Graph-CNN would possess a graph signal with $\mathbb{R}^N$ vertices per feature map, with an increasing number of output maps leading to scaling inefficiencies without a pooling operation. Such a graph coarsening strategy looks to remove or aggregate vertices in a given graph, while retaining key information in the graph structure. The coarsening of $G = \{V, W\}$ into a reduced graph, $\hat{G} = \{\hat{V}, \hat{W}\}$, is non-trivial, with a wide variety of methods into reducing graph and signal complexity, Liu et al. (2014), Safro et al. (2012), Safro (2009), Shuman et al. (2016) and Ying et al. (2018). Since the graph convolution via the GFT requires a fixed Fourier basis, it is

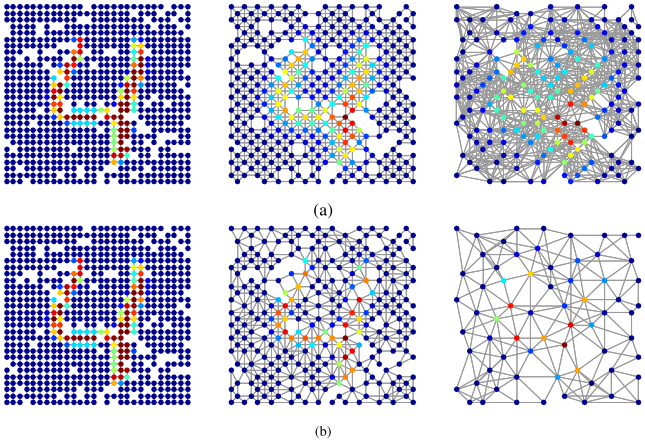


Fig. 4. Two levels of graph pooling operation on an irregularly sampled 2D grid. (a) Kron's reduction, (b) AMG. Note that both methods retain overall spatial structural distribution, however the edge connectivity of Kron's reduction results in an explosion in edge count.

possible to pre-compute the required graphs for the architecture before training, and look them up for convenience. Two methods for graph coarsening are the use of AMG and Kron's reduction pyramids (Safro, 2009; Shuman et al., 2016), a comparison of which is shown on an example of a irregularly sampled 2D grid in Fig. 4. Selecting a collection of vertices to keep in the coarsened graph can take several forms, including a selection criteria based on the polarity of the eigenvector associated with the largest eigenvalue, $\hat{V} = \{U_{N,i}\}; U_{N,i} \geq 0$, or the use of spectral clustering of the vertices via k-medoids over the eigenvectors (Liu et al., 2014; Safro et al., 2012). Once we have defined the graph-based convolutional network operators we are able to construct a Graph-CNN architecture as we would with a CNN, with initial feature mining layers and a subsequent fully connected head. Our proposal utilizes the outlined multi-resolution sampling operation and the use of a Graph-CNN driven MSL approach, upon which we will provide a case study in medical volume structure segmentation.

3. Case application: 3D medical image segmentation

The use of multi-scale learning is beneficial in many domains, and the proposed usage of multi-resolution sampling and deep learning on the irregular domain is generalizable to the overarching problem of multi-scale representation learning. In this study we provide evaluation of the proposed multi-resolution Graph-CNN on the domain application of medical segmentation, in which the understanding of detailed localized information and the general wider context of the human anatomy is beneficial in detecting small-scale anatomical structures with accuracy (Li et al., 2017; Yan et al., 2017; Gao et al., 2014; Chen et al., 2015; Zhang et al., 2017). Recently, there has been tremendous work in the application of neural network methods to medical image analysis (Jiang et al., 2010), and in particular CNNs for anatomical organ detection (Shin et al., 2016; Kamnitsas et al., 2017) and unsupervised learning (Shin et al., 2013). Segmentation is a key area in image analysis and many applications make use of segmentation methods to process a volume into meaningful parts, especially medical volume analysis. Such methods often attempt to label each voxel of a volume into a given class of interest, utilizing appearance information or some structural features extracted from the volume. One such application of medical volume understanding is the segmentation of the aortic valve. Aortic valve stenosis is a common heart disease affecting 3% of the global population, with many cases requiring surgical treatment and the 3D segmentation of the aortic root is beneficial for patient selection, procedural planning and post-evaluation. It is therefore vital to reliably and accurately identify aortic root structure within a patient, Fig. 5.

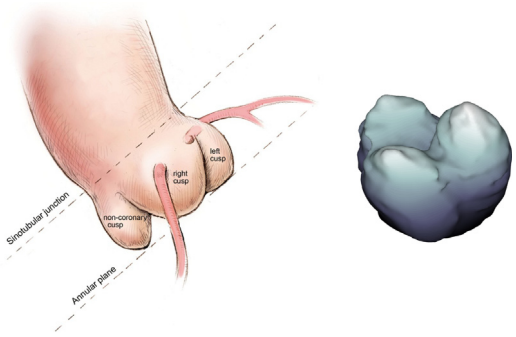


Fig. 5. Structure of the aortic valve. Left: Diagrammatic representation of the aortic root and valve, detailing the three cusps. Right: Ground truth mesh from the dataset, oriented vertically. Note that ground truth is labeled up to the sinotubular junction. *Source:* Image from Cheung and Lichtenstein (2012).

Due to image noise and other ambiguities, non-model based approaches are often unable to detect subtle boundaries between classes in a volume, e.g. those between the valve and left ventricular output tract (Zheng et al., 2012). However, given an initial shape, deformable models are able to identify this boundary, and have successfully been used for segmentation of the root structure (Zheng et al., 2012; Grbić et al., 2012; Palmer et al., 2015). Structure generalization and application of priors are often key in methods that perform detection and segmentation of medical imaging data.

Supervised automatic 3D deformable modeling is not only computationally demanding during the testing stage, but it is also labor intensive in preparing training data, e.g. in establishing correspondence for smooth 3D structures. Parameters for effective model regularization as well as useful feature extraction are chosen carefully depending on the application, which can be extremely time consuming. For example, model regularization regularly requires building a statistical model which often demands additional manual labeling (Cootes et al., 1995). Similarly, choosing optimized discriminative features for both object and boundary detection can be an excessively lengthy process. In this work, we aim to alleviate the burden of feature crafting, as well as implementing an efficient segmentation method using a bottom-up approach with prior regularization.

Appropriate automatic solutions to building statistical models are not well reported in the literature. Notably however, Frangi et al. (2002) proposed finding mesh correspondences based on image data rather than the meshes themselves. The meshes were locally transformed to an atlas and anatomical points were propagated across the set. The transformation was estimated with intensity-based mutual information which is not suitable for noisy images with relatively low contrast between soft tissues, such as cardiac CT data. As such, we propose estimating the transformation using a mesh-based similarity metric and learned correspondences between training samples. The proposed method eliminates the process of manual landmark labeling, enabling a larger set of fiducial points per shape and providing a reduction in overall time taken to construct a shape model.

To initialize deformable models, they need to first be automatically aligned with the test image by performing object detection. Exhaustive pose parameter search in 3D is highly impractical due to possible position, orientation and scale permutations. Alternatively, MSL has been proposed for efficient 3D organ detection (Zheng et al., 2012; Grbić et al., 2012; Palmer et al., 2015) by incrementally searching position, position-orientation and position-orientation-scale spaces. Zheng et al. (2009) presents a method for further reducing the parameter space by further constraining the initial search space based on the statistical correlation between pose parameters in the observed training data, further removing the testing of unnecessary hypotheses. Choosing the appropriate features for classifiers is challenging, as feature type, orientation and scale must be considered, and pathological structures

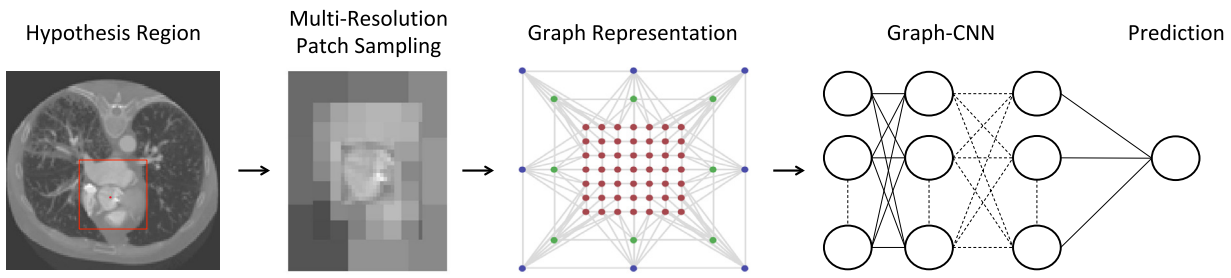


Fig. 6. Illustration of proposed marginal space learning classifier. Multi-resolution pooling reduces a large patch into a contextually and locally informative graph signal for use in a Graph-CNN architecture. A Graph-CNN architecture is then constructed using graph convolution and pooling layers to obtain a prediction on position or position-orientation of the observed patch. Color of nodes within the graph representation indicates resolution of a given layer, i.e. lowest outer resolution sampling to the volume edge (blue), highest resolution in central core (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

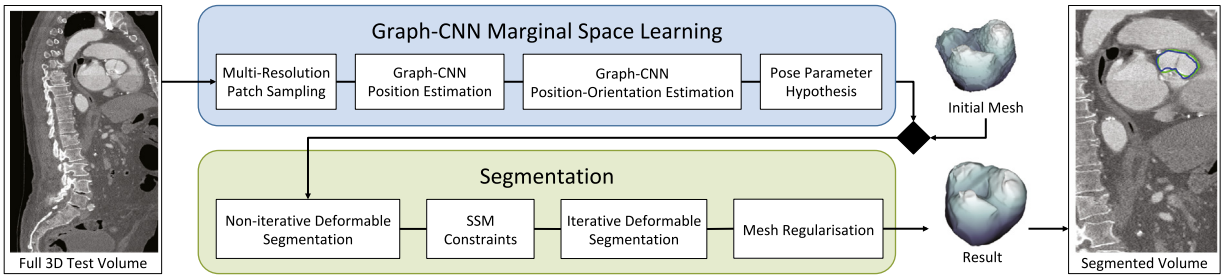


Fig. 7. Our testing pipeline for the proposed segmentation. Output segmentation magnified for clarity.

often look significantly different between observations. MSL introduced by Zheng et al. (2008) and Zheng et al. (2012) provides a hand-crafted steerable feature extractor which is used as to produce input for the pose estimation classifiers. We argue that a feature learning based approach should be adopted, such as those obtained through deep learning architectures. The mining of such features has shown marked improvements on the start of the art in numerous image analysis problems (LeCun et al., 2015), and we further propose that incorporating a sense of spatial context into the learning strategy can improve the accuracy of the classifier model produced for MSL.

The development of an end-to-end segmentation pipelines have only seen very recent study (Long et al., 2014; Caesar et al., 2016), and even fewer have been applied to volumetric data due to the complexity of dense segmentation of large volumes. Numerous dense segmentation methods exists, utilizing the fully convolutional neural network approach to provide a segmentation of input images (Xue et al., 2017). Milletari et al. (2016) go one stage further and present a fully convolutional neural network for volumetric segmentation of medical images, providing a dense segmentation model which takes 48 h to train on the 128 × 128 × 64 volumes. This has since been expanded by Zhang et al. (2018) to utilize the Region Proposal Network structure from the Faster R-CNN approach to localize and focus the convolutional attention (Ren et al., 2016). Both methods are dependent on the use of a fixed sampling resolution, limiting the ability to consider local and contextual features without increasing complexity with a branched multi-scale network architecture.

Irregular sampling within MSL has been approached before, notably by Ghesu et al. (2016), where a sparse sampling is taken from within a localized area of fixed resolution and fed to a fully-connected neural network architecture to drive the pose estimation classifiers. The use of a fully-connected network removes the intrinsic spatial relationships between the input features, whereas the proposed method differs from this approach by introducing the use of a patch sampling from across multiple resolutions and by incorporating spatial relationships between the input features by training a Graph-CNN architecture. The proposed multi-resolution sampling provides local and contextual information to the network during training. We also introduce an automated shape model landmark detection approach, providing data-driven statistical shape model generation and reducing user input.

Overall, we present a novel pipeline method of deep learning on the graph representation of an irregular multi-resolution spatial domain for identifying target position and orientation hypotheses in aortic root detection. Raw local and contextual intensity features are used in a novel Graph-CNN architecture to mine spatially related features on an irregular spatial topology, avoiding relying upon hand-crafted features or an increased overhead from large patches. A marginal space learning approach is taken to reduce the search space complexity of the large 3D parameter space for segmentation initialization. An initial shape model is learned in an automated fashion by detecting a set of landmark features across the training meshes; reducing the manual effort of labeling fiducial landmarks on each mesh and allowing for a larger set of landmarks to be identified. A deformable segmentation framework is proposed that does not rely heavily on top-down constraints, instead presenting a non-iterative deformation and shape model regularization step for the initial segmentation of the volume. This is then followed by an iterative refinement of the mesh with local deformations and mesh-based regularization based on a strong boundary detection network. The use of SSM shape constraints and mesh regularization utilizes prior information regarding learned shape context in order to produce a data-driven segmentation with reduced mesh entanglement and user guidance. Results on the proposed method show strong performance benefits in both aortic root pose estimation for the purpose of marginal space learning, and an accurate segmentation of the aortic root structure. Evaluation of the proposed approach is given in the medical segmentation domain.

4. Proposed approach

The proposed method looks to improve the efficiency of the overall volume processing pipeline in an automated fashion, removing the need to manually label landmark points on numerous training samples by hand, introduce multi-scale learning into the MSL pipeline and to utilize constrained deformable segmentation to produce regularized meshes. Our multi-resolution graph-based sampling produces a patch without overlapping regions, reducing the number of elements sampled

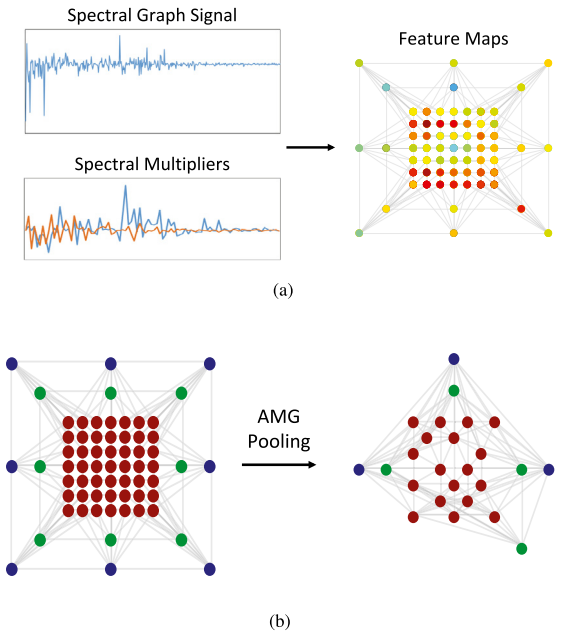


Fig. 8. Graph convolutional neural network operators. (a) Graph Convolution. Spectral graph signals are multiplied with spectral multiplier filter weights. An inverse Graph Fourier Transform returns the signal to the spatial domain. (b) AMG Pooling, fine-scale nodes are aggregated into coarser nodes in the pooled graph. Color of nodes indicate resolution of a given layer, i.e. lowest outer resolution sampling to the volume edge (blue), highest resolution in central core (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in comparison to the equivalent area covered with a conventional regularly sampled patch. Through the use of multi-resolution sampling and Graph-CNN we are able to use a local and contextual sampling patch to feed the network whilst learning localized filters on the multi-resolution graph via graph signal processing operators, Fig. 6.

The proposed testing framework is shown within Fig. 7. Localization and alignment parameters for the initial mesh, a median mesh from the training set, is efficiently carried out using a novel Graph-CNN-Based Marginal Space Learning (Graph-CNN-MSL) approach. Deformable segmentation is composed of boundary detection and 3D mesh regularization allowing large-scale movements by setting long search paths at the boundary detector stage, with shape constraints applied to avoid mesh and shape irregularities. Local refinement is then performed using deformable segmentation in an iterative fashion, where each iteration is capable of small movements, followed by generic mesh processing.

5. Methodology

The proposed method consists of two major parts; the use of Graph-CNN-MSL to reduce the complexity of parameter search space, and SSM supported segmentation to generate an accurate and regularized mesh of the aortic root, Fig. 7. The following section outlines the proposed pipeline components.

5.1. Marginal space learning

We train two classifiers which act as detectors for a positive position and orientation within the search space. For our MSL estimators we define the search space criteria as outlined in Section 5.1. The position estimator is composed of search parameters $P = (x, y, z)$, $P^t = (x_t, y_t, z_t)$, $S = (1, 1, 1)$ voxels, and the input layer features are the intensities from our globally aligned pooling. For the position-orientation estimator $P = (x, y, z, \omega, \phi, \theta)$, $P^t = (x_t, y_t, z_t, \omega_t, \phi_t, \theta_t)$, $S = (1, 1, 1, 10°, 10°, 10°)$, and the patches are aligned with the orientation hypothesis. An example volume, with 512×512×512 voxels and full 360° orientation space about the X, Y, and Z axes, would result in over 6.26×10^{15} pose parameter hypotheses. MSL allows us to first search 1.34×10^8 position parameters, select the top 10 probable position hypotheses, and then search roughly 4.66×10^8 position-orientation parameters. This is an overall reduction on the order of 10^7 over exhaustive search of the pose space. Finally, for simplicity, we use the mean local scale of the training meshes to yield a 9-element pose estimation vector $(x, y, z, \omega, \phi, \theta, S_x, S_y, S_z)$. The use of mean scale incorporates the base scale information used in the multi-resolution sampling, simplifying the process over creating an appropriate scale search space.

5.2. Multi-resolution graph-CNN

To produce a graph representation of the multi-resolution sampling operation, for each selected resolution level l, we generate a set of Cartesian coordinates, P^l, sampled at the given resolution rate. We then remove points from P^l that are spanned by $P^{1:l-1}$, discarding observed regions of overlap. Intra-layer edge weighting is calculated as

$$w_{(i,j)} = e^{-\frac{\|v_i - v_j\|^2}{\sigma}} \tag{6}$$

on an epsilon nearest neighborhood of vertex v_i, with a search radius of $\epsilon = l$, the current sampling resolution, where $\|v_i - v_j\|$ is the L^2 distance between the vertices v_i and v_j and $\sigma = \frac{\epsilon^2}{2}$. This returns the 4-way Von Neumann neighborhood relationships of the vertices within a resolution layer. Inter-layer edges connect the lower-resolution layer vertices to the higher-resolution core via the l nearest neighbor vertices, relating wider contextual features to the high-resolution region of interest within the patch. Weighting for inter-layer edges are defined by scaling Eq. (6) down by the current resolution factor, with v_i representing a vertex in the current layer, and v_j a vertex in the high-resolution core.

For this application the selected graph pooling operation is an AMG graph coarsening, selecting vertices in the finer graph resolution for aggregation into coarser vertices within the pooled graph and avoiding an explosion of edges in the coarsened graph when compared to the use of Kron's reduction (Safro, 2009), Fig. 4. Aggregation takes spatially localized subsets of V from G, and generates a singular vertex for each subset in the new set of vertices $\hat{V}$ in the output graph $\hat{G}$. The graph signal, $f_{1:N}$, associated with G is then down-sampled to reside on $\hat{G}$ as the coarser graph signal $\hat{f}_{1:n}$, where N and n are the original number of vertices and the pooled number of vertices respectively. The AMG operation produces a set of projection matrices, restriction matrix R and interpolation matrix I, for down-sampling and up-sampling transform of the graph signals between the finer and coarser graph scales. To coarsen the graph signal f^0 from residing on G^0 to f^1 on the coarsened graph G^1 we utilize the restriction matrix R^0 by

$$f^1 = f^0 R^0 \tag{7}$$

and to reconstruct the signal we can interpolate through I^1 as

$$f^0 = f^1 I^1 \tag{8}$$

The Graph-CNN models are built as follows. (1) A graph representation of the multi-resolution volume space is generated; (2) Grayscale voxel intensities from the volume are samples using the multi-resolution sampling scheme; (3) Multi-resolution pooling yields a significantly reduced representation on the irregular graph; (4) Pooled patch values are fed into the Graph-CNN, undergoing graph spectral convolutions and graph pooling operators, Fig. 8; (5) An output detection prediction is given for each observed hypothesis.

5.3. Statistical shape model

The proposed segmentation stage first uses the predicted hypothesis pose parameter vector $(x, y, z, \omega, \phi, \theta, S_x, S_y, S_z)$ to align the initial shape model, a median shape from the training set, to the volume. A boundary detector is then used to guide a non-iterative local deformation that is

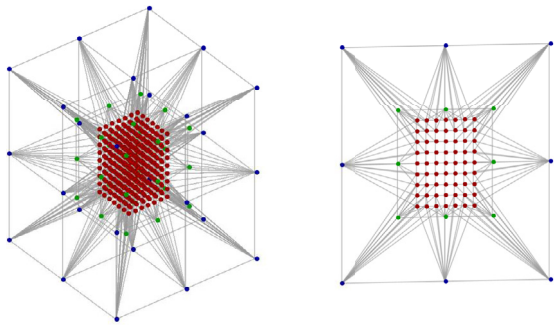


Fig. 9. (a) 3D multi-resolution volume graph for orientation estimation. (b) 2D example, note removed nodes in regions of overlap. The high-resolution core is a cuboid structure, extending along the Z-axis to capture further information about the ascending aorta. Node coloring represents variable resolution sampling, from low resolution outer layer to the high resolution inner core. Best viewed in color.

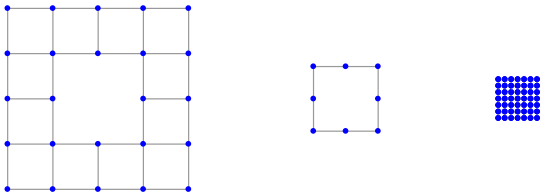


Fig. 10. Exploded view of an example 2D multi-resolution graph. Note the empty center of each successive outer layer, and the irregular sampling distances between layers. These combining factors make such a multi-resolution sampling domain unsuitable for current convolutional neural network approaches.

then constrained via a shape regularization step. Mesh refinement is then obtained via an iterative application of local deformations.

To automatically label training data and identify a SSM for deformation regularization we propose locally transforming meshes to a reference mesh and propagating points across the set. The automated landmark detection allows for a larger set of landmarks to be identified with little impact on the user. To construct the initial shape model, a target mesh $M_t = (V_t, E_t, F_t)$ with $|V_t| = n$ vertices is randomly selected from the training set, and a subset of m fiducial point vertices are labeled such that $P_t \subset V_t$, and $m \ll n$. All other meshes in the set are regarded as source meshes, such that $M_s = (V_s, E_s, F_s)$ where $|V_s| = p$, and $n \neq p$. The aim is to find a subset of m vertices $P_s \subset V_s$, that are correspondent with P_t. We work on the assumption that finding correspondences between two shapes becomes much easier if the shapes are similar to each other. Therefore we apply a transformation $T(x, y, z) : M_s \mapsto M_t$, consisting of global $T_g(x, y, z)$ and local $T_l(x, y, z)$ transformations.

T_g globally aligns both meshes and is formulated as an affine transformation from ground truth vectors. T_l then takes into account the local differences in shape, and is estimated using (10), (11), and a similarity metric

$$E(V_t, V_s') = \sqrt{\sum_{i}^{n} (V_t - V_s'')^2}\,, \tag{9}$$

where V_s'' are the corresponding nearest neighbor vertices in V_s' with respect to V_t. For every point in P_t, its nearest neighbor based on Euclidean distance is found in V_s', resulting in P_s'. Finally, applying T_l^{-1} to P_s' yields P_s.

5.4. Deformable segmentation

After alignment of the initial mesh with pose parameters identified by MSL, the non-iterative deformation stage utilizes appearance features to adjust the vertex set by defining a search path along the normal direction. A boundary detector is trained to find the path coordinate with the optimal boundary response, and landmark vertices are deformed to these positions. In order to avoid hand-crafting features, we again utilize feature learning. For boundary detection we use a shallow fully-connected Neural Network (NN), learning low-level features from a small set of intensities on a local patch, centered at the search path coordinate and aligned with the path direction. The small area of observation ensures iterative refinement of the mesh is based upon response to localized boundary features, with little interference from wider appearance. A 3×3 patch is extracted from each point along a boundary search path, vectorized, and input to a single-layer neural network.

Boundary detection now results in new vertex positions V', however there is potential for mesh entanglement amongst the new set of vertices. To counter this we use B-spline based 3D mesh regularization, Palmer et al. (2015), where a non-rigid transformation $T(x, y, z)$ is estimated between V and V' before performing a free-form-deformation (FFD) on V to fit V'. To estimate $T(x, y, z)$, control points $\phi^r_{i,j,k}$ separated by δ, are moved which warp an underlying 3D voxel lattice. Given a set of control points, the transformation is formulated as follows:

$$T(x, y, z) = \sum_{l=0}^{3}\sum_{m=0}^{3}\sum_{n=0}^{3} B_l(u) B_m(v) B_n(w) \phi_{i+l,j+m,k+n}\,, \tag{10}$$

where B_l represents the lth basis function of the B-spline, $[i, j, k]$ are the voxel positions, and $[u, v, w]$ are the fractional positions. The positions of $\phi^r_{i,j,k}$ are optimized using gradient descent consisting of a smoothness cost and a sum-of-squared-difference similarity metric between V (after warping) and V'. The final transformation is estimated at multiple resolutions R, similar to FFD registration (Rueckert et al., 1999):

$$T^R(x, y, z) = \sum_{r=1}^{R} T^r(x, y, z)\,. \tag{11}$$

For our purpose, $R = 3$, and at each mesh resolution the control point spacing is $\delta_r = \delta_0/2^r$, which controls the FFD degrees-of-freedom. After applying SSM constraints during segmentation, only corresponding fiducial points are regularized. Thin plate spline warping is used to interpolate remaining vertices, resulting in ~8000 final corresponding vertices.

The next stage of the pipeline is to take the regularized mesh and perform iterative refinement of the mesh boundary by applying repeated rounds of mesh deformation with the NN boundary detector. This avoids a heavy top-down constraint on the segmentation, instead having a single round of top-down shape constraint followed by an iterative data-driven refinement stage. Vertices are iteratively deformed by identifying boundaries along the normal direction as above. A final generic mesh-processing step rounds out the pipeline to regularize and smooth the mesh for output.

6. Implementation

For evaluation of fully automated Graph-CNN-MSL and segmentation, we provide an example application on aortic valve segmentation from 3D-CT scans. We perform 3-fold cross-validation via segmentation on 36 3D-CT volumes of size $512 \times 512 \times (500 \sim 800)$ and voxel size $0.48\ \text{mm} \times 0.48\ \text{mm} \times 0.62\ \text{mm}$. Benefits of utilizing Graph-CNN architectures to learn spatially related features for MSL are evaluated in comparison to use of standard NN and hand-crafted feature classifiers. Comparison of the proposed segmentation stage is given against a state-of-the-art method and traditional statistical shape model based segmentation.

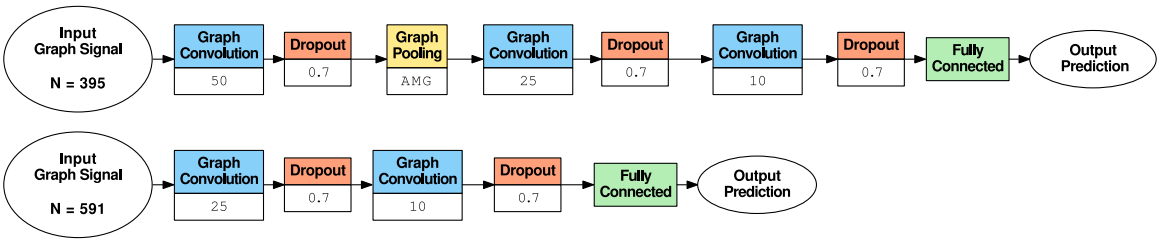


Fig. 11. Graph-CNN MSL network architectures. Top: Graph-CNN position estimator. Bottom: Graph-CNN orientation estimator. Input graph signals result from the multi-resolution patch sampling outlined in Section 6.1.

6.1. Marginal space learning

Our multi-resolution position estimator consists of a patch comprised of 3 resolutions; an inner core of $1 \times 1 \times 1$, a middle region of $2 \times 2 \times 2$, and an outer region of $3 \times 3 \times 3$ times the mean local scale. Multi-resolution layers were pooled at resolutions of $\frac{1}{8}$, $\frac{1}{40}$, and $\frac{1}{56}$, resulting in a graph with 395 vertices. Position-orientation inputs were taken from a patch at 3 resolutions; an inner core of $1 \times 1 \times 1.2$, a middle region of $2 \times 2 \times 2$, and an outer region of $4 \times 4 \times 4$ times the mean local scale. Regions were pooled at resolutions of $\frac{1}{8}$, $\frac{1}{40}$, and $\frac{1}{56}$ respectively, resulting in a graph with 591 vertices. The inner high-resolution core of this patch is cuboid in shape, extending along the Z-axis to provide further high detail information about the ascending aorta to assist with orientation estimation. Coordinates generated from this multi-resolution setup were then used to generate the graph structure for the Graph-CNN operators as defined in Section 5. Fig. 9 shows the resulting graph structure for both classifiers utilized in Graph-CNN-MSL, whilst Fig. 10 outlines the intra-layer connectivity for each of the sampled resolutions. The graph signals residing on this graph representation correspond to the voxel intensities sampled via the multi-resolution sampling scheme, with the values at each node aligned to the grayscale intensity of the original image sampled at that node's respective resolution and location.

Two separate Graph-CNN architectures, Fig. 11, are utilized to estimate position and position-orientation parameters for shape model alignment. The aortic root position estimation architecture was empirically defined as $C^{50}PC^{25}C^{10}$, where C^o is a graph convolutional layer with o output feature maps and P is an AMG graph pooling layer. Each graph convolutional layer is followed by Rectified Linear Unit (ReLU) activation, batch normalization, and dropout to further support generalization of features and reduce model overfitting. A binary classification outputs predicted labels and provides a back-propagation target for training. Networks were trained using an ADAGRAD optimization strategy (Duchi et al., 2011), with an initial learning rate of 10^{-3} and batch size of 32. Training samples were selected with an object/non-object ratio of $\frac{1}{50}$. The orientation estimator architecture was $C^{25}C^{10}$, and utilized an object/non-object ratio of $\frac{1}{25}$.

In order to identify the benefit of utilizing a localized feature extraction constraint provided by the Graph-CNN architecture, a fully-connected neural network was constructed where C^o layers are replaced with fully-connected layers of the same size as in architectures above. These fully-connected networks had no intrinsic representation of spatial relationships between features, essentially representing a fully-connected and equally edge-weighted graph, as represented in Fig. 12. Training hyper-parameters of the neural network implementations were kept the same as with the Graph-CNN, utilizing ADAGRAD optimization with a learning rate of 10^{-3} and batch size of 32.

Our boundary detector was trained with an equal boundary/non-boundary ratio using intensities from a 3×3 local patch. Patches were fed through shallow fully-connected network in order to learn low-level boundary features. A comparison hand-crafted feature based approach utilized steerable features and a boosted tree ensemble classifier, as per (Zheng et al., 2012).

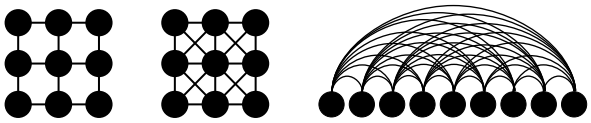


Fig. 12. Examples of graph constructions for a 3×3 regularly spaced grid structure. Left to right: Von Neumann Neighborhoods (4 way), Moore Neighborhoods (8-way), fully-connected (non-spatial if equally weighted). Choice of a suitable graph construction approach is required as the graph represents underlying spatial relationships within the domain.

Table 1
Predictive accuracy of the marginal space learning approaches. The addition of a locally receptive filtering operation within the proposed Graph-CNN approach provides an improvement over the standard fully-connected neural network method, lowering both the position and the orientation error of the predicted pose parameters.

MSL method	Position (Voxels ± SEM)	Orientation (Degrees ± SEM)
Hand-crafted	9.10 ± 0.57	14.69 ± 1.28
Fully-connected	3.79 ± 0.47	12.38 ± 1.24
Proposed	$\mathbf{1.46 \pm 0.36}$	$\mathbf{6.78 \pm 1.01}$

6.2. Segmentation

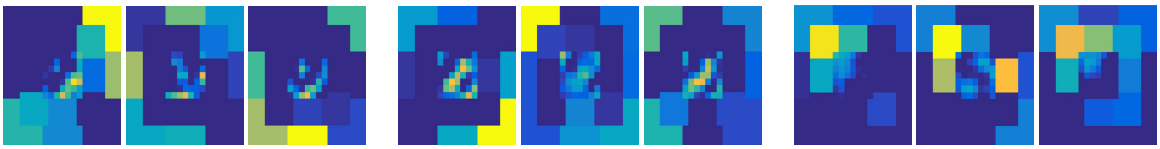
To generate the required initial shape model for deformable segmentation, we label 70 fiducial points on a single target mesh, which were propagated across the remaining training set as set out in Section 5. We compared the proposed segmentation pipeline with two competing methods; a modified Active-Shape-Modeling (ASM) implementation, and another deformable modeling method (Zheng et al., 2012). Zheng et al. (2012) consisted of a boundary detector trained with steerable features, followed by Taubin mesh smoothing in an iterative fashion for mesh refinement. For fair comparison, we included a 3D mesh regularization stage in our implementation of ASM.

7. Comparative analysis and results

We report evaluation on both MSL and deformable segmentation portions of the pipeline, outlining contribution of Graph-CNN feature learning for aortic root position and orientation parameter estimation, and non-iterative shape deformation with regularization for segmentation.

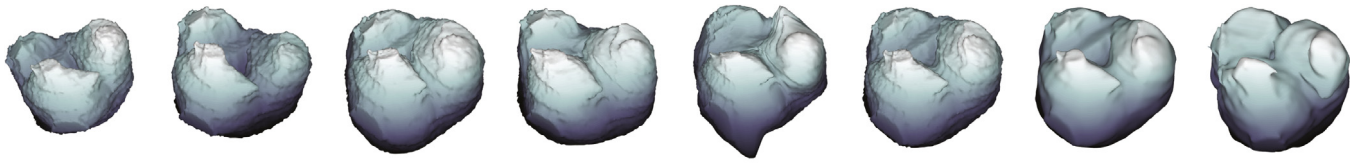
7.1. Marginal space learning

A comparison of classifier methods utilized for MSL is presented in Table 1. The self-learning feature mining methods of NN and Graph-CNN outperform use of hand-crafted features for both position and position-orientation estimation, with Graph-CNN further improving over NN architecture. Being able to accurately and reliably provide hypothesis regions on which to initialize a segmentation algorithm is highly beneficial to following segmentation steps. The Graph-CNN position detector's sensitivity and specificity were 91.46% and 99.95%,



(a) (b) (c)

Fig. 13. Example feature maps from positive patches from 3 separate test volumes. Feature-maps from filter responses to (a) local and contextual features, (b) local features (c) non-local features.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 14. Output at each stage of segmentation. (a) Initial shape model, (b) pose alignment, (c) non-iterative deformation, (d) SSM constraint, (e) iterative deformation, (f) final mesh regularization, (g) post-segmentation smooth, (h) ground truth.

respectively. Sensitivity and specificity of the position-orientation estimator was 89.84% and 84.16%. Given the huge parameter search space, strong specificity results are invaluable to reliably reduce parameter search spaces and greatly increase efficiency. The proposed method provides a significant increase in accuracy over both the hand-crafted and fully-connected neural network implementations. Results showed a significant difference in position estimation accuracy between Graph-CNN and hand-crafted features, $t(35) = -11.76$, $p < .001$. Graph-CNN also provides benefit in orientation estimation over the hand-crafted feature approach, $t(35) = -9.37$, $p < .001$. The NN method outperformed the hand-crafted feature approach, as shown in Table 1; the trained estimators provided a significant improvement over hand-crafted features for both position and orientation, $t(35) = -7.74$, $p < 0.001$ and $t(35) = -7.35$, $p < 0.001$ respectively. When comparing the spatially localized feature learning of the Graph-CNN architecture against the fully-connected neural network approach, Table 1 shows that both the position and orientation estimator see marked improvements, with $t(35) = -4.31$, $p < .001$ and $t(35) = -4.89$, $p < .001$ respectively.

The Graph-CNN architectures provide a large gain in accuracy over the other methods, utilizing spatial relationships between the high and low resolution spaces in a single descriptor. One benefit of utilizing the underlying spatial topology of the problem domain is the ability to visualize the learned descriptors (Zeiler and Fergus, 2014; Yosinski et al., 2015), and the same can be utilized in Graph-CNN methodologies to identify the localized responses within the network's filters. Fig. 13 shows some example feature maps produced by graph convolutions of spectral filters in the Graph-CNN model trained for the position classifier. The feature maps describe the activations of 3 filters (a, b, c) on multi-resolution patches centered at the ground truth of 3 different volumes to show the response to the aortic root structure. The visualization plots a slice through the center of the multi-resolution volume with the topology as in Fig. 9. From these visualizations we are able to identify activation responses to the local and contextual information within the multi-resolution patch, with the outer layers responding to lower resolution features from the wider contextual region surrounding the patch center. Fig. 13a shows consistent responses across high and low resolutions. Fig. 13b displays features which are consistently found in the high resolution region, in this example a diagonalized response across the center. Fig. 13c shows features found in the mid-resolution layer of the graph.

7.2. Segmentation

Segmentation performance is evaluated in terms of the symmetrical point-to-mesh error and symmetrical Hausdorff distance. Mesh error provides an indication of average error in the predicted segmentation, however the Hausdorff distance gives insight into the presence of outlying regions on the predicted mesh. Overall the proposed pipeline showed notable improvements in segmentation accuracy compared to the comparison methods, with an average Mesh Error of 1.27 ± 0.23 mm and a Symmetrical Hausdorff Distance of 6.04 ± 1.50 mm (Table 2). The benefit of regularization for suppressing mesh entanglement can be seen by the lower symmetrical Hausdorff distances of the regularized ASM and proposed methods. The use of deformable segmentation refinement helps to drive the mesh error lower, iteratively bringing points closer to the appearance boundaries identified by the shallow network. For both error metrics the proposed method shows lower standard deviation, with the pipeline providing consistently accurate and reliable segmentation. The proposed method provides a significant improvement over the ASM approach for both mesh error and Hausdorff distance, with $t(35) = -7.17$, $p < .001$ and $t(35) = -3.45$, $p = .0015$ respectively. Compared to the method provided by Zheng, only the Hausdorff provided a significant difference in performance with $t(35) = -7.68$, $p < .001$. There was no significant difference between the proposed method and that of Zheng in regards to their mesh error, with $t(35) = 1.25$ and $p = .22$

Table 2
Comparison of segmentation approach accuracy.

Method	Mesh error (mm ± STD)	Hausdorff distance (mm ± STD)
Regularized ASM	2.01 ± 0.63	9.13 ± 2.58
Unregularized deformation	1.44 ± 0.59	10.29 ± 2.93
Proposed	$\mathbf{1.27 \pm 0.23}$	$\mathbf{6.04 \pm 1.50}$

Table 3 highlights findings comparing the overall pipelines. First, proposed use of Graph-CNN for mesh pose initialization provides a consistent benefit to the segmentation portion of our pipeline. Second, proposed segmentation steps are able to produce meshes with low Hausdorff Distance to the ground truth, a benefit of regularization for controlling mesh entanglement. It can also be seen that Graph-CNN methods provide low standard deviation across numerous test volumes, indicating that accurate pose parameter hypotheses from Graph-CNN-MSL are beneficial to the following segmentation phase.

7.3. Qualitative segmentation analysis

Output from each stage of the segmentation pipeline can be seen in Fig. 14, detailing the alignment of an initial mesh to pose parameters

Table 3
Comparison of MSL initialization methods on final segmentation performance.

Segmentation method	MSL method	Mesh error (mm ± STD)	Hausdorff distance (mm ± STD)
Regularized ASM	Hand-crafted	2.01 ± 0.63	9.13 ± 2.58
	NN	2.00 ± 0.78	8.42 ± 3.28
	Graph-CNN	**1.66 ± 0.45**	**6.92 ± 2.05**
Unregularized deformation	Hand-crafted	1.44 ± 0.59	10.29 ± 2.93
	NN	1.51 ± 0.66	10.59 ± 3.41
	Graph-CNN	**1.23 ± 0.27**	**9.10 ± 2.26**
Proposed	Hand-crafted	1.50 ± 0.56	7.72 ± 3.20
	NN	1.49 ± 0.52	7.85 ± 3.24
	Graph-CNN	**1.27 ± 0.23**	**6.04 ± 1.50**

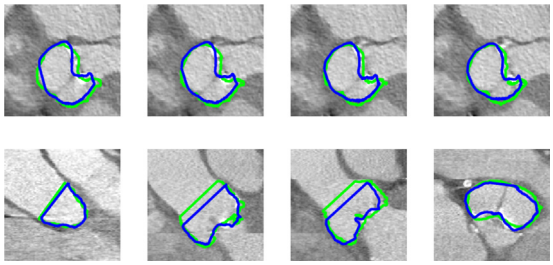


Fig. 15. Segmentation shown on cropped image slices for illustration. Green contours show ground truth, blue contours show result of proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

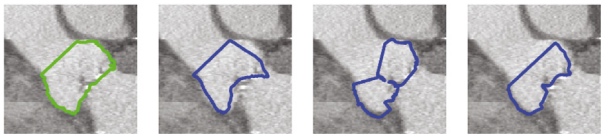


Fig. 16. Segmentation comparison for three pipeline methods. (a) Ground Truth, (b) Modified ASM, (c) Zheng Zheng et al. (2012), (d) Proposed.

identified via Graph-CNN-MSL, non-iterative deformation, application of the SSM constraint, and the iterative deformation stage. Given the difference in pose between ground truth and median initial shape, it is important to identify accurate pose parameters for shape alignment.

Final segmentations and their cropped slices are shown in Fig. 15, including different axial views, and Fig. 16 compares segmented slices from each evaluated method. Entanglement is observed in slices implementing a top-down approach with no regularization (Zheng et al., 2012), Fig. 16c, whilst the modified ASM fails to expand and meet the boundary contour due to the heavy shape constraint. The proposed method shows it is able to maintain a smooth regularized mesh whilst also deforming towards the boundaries in all spatial directions.

Although the point-to-mesh error of Zheng et al. (2012) is marginally lower (~0.04 mm) than the proposed method when initialized with Graph-CNN-MSL, these meshes lack regularization, resulting in the higher symmetrical Hausdorff distance error. It is also worth noting that our method is automated at the training stage, whereas (Zheng et al., 2012) requires extensive manual preparation time to produce suitable hand-crafted feature vector representations and identify landmark points across training meshes. By labeling landmark points in an automated fashion we are able to greatly reduce the pre-processing time required to start model training. Local transformation to identify a corresponding subset of vertices across training meshes allows scaling of identified fiducial landmark points in our shape model without drastic increase in pre-processing effort, as seen by the proposal to identify 70 landmark points compared to the 8 within Zheng et al. (2012).

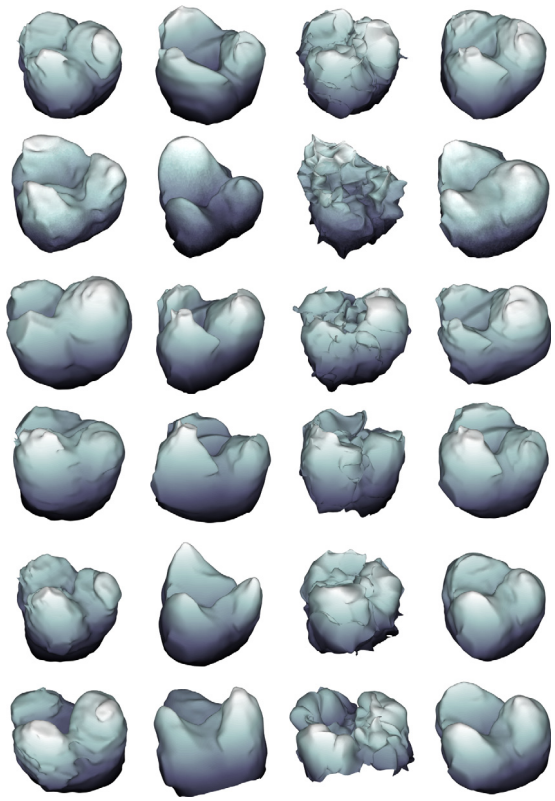


Fig. 17. Segmentation comparison, where each row is the result of a different test image. Columns: (1) Ground truth; (2) modified ASM; (3) Zheng et al. (2012); (4) Proposed method.

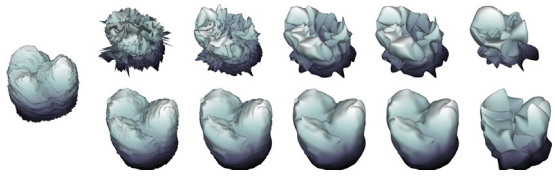


Fig. 18. Repeated Taubin smoothing of meshes. Starting with an unregularized mesh (top) and a regularized mesh (bottom). Left to right: Ground truth, Applications of smoothing (iteration): 0, 1, 10, 20, 200. The entanglement of the mesh remains through the application of smoothing and the excessive smoothing results in meshes eventually beginning to diverge from the ground truth. Note that ground truth labeling is not locally smooth due to labeling process.

Mesh comparisons in Fig. 17 show that some shape constraint is beneficial for generating ordered mesh surfaces. The meshes produced by Zheng et al. (2012) are significantly entangled compared to both the proposed method and the modified ASM, however the modified ASM produced high point-to-mesh errors due to the lack of shape deformation towards the structural boundaries. This shows that strong shape generalization can be too restrictive, and it is critical not to overly rely on top-down constraints. We applied the Taubin smooth as a final mesh smoothing operation to both our proposed method and the comparison method from Zheng. We have also explored the effect of increasing the smoothing effect on the predicted meshes. As can be seen in Fig. 18, the repeated smoothing does not correct the mesh entanglement but can initially reduce surface variance. The observation here show that reliance on smoothing as a method for repairing the mesh surface is not an optimal approach, and instead an integrated mesh regularization approach which avoids mesh entanglement provides an initial prediction of a well-structured mesh surface which then be improved slightly with smoothing.

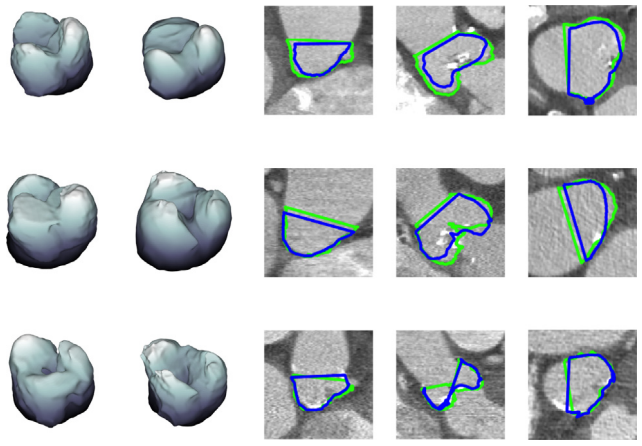


Fig. 19. Example failure cases of the proposed pipeline. Left to right: Ground truth mesh, predicted mesh, yz-slice, xz-slice and xy-slice through segmented volume. Green contours show ground truth, blue contours show result of proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Failure cases are shown in Fig. 19. Visualization of the predicted mesh shows that the overall shape is often reasonable, with the overall model shape, including root and cusps, being present and well formed, however there are some issues with orientation alignment (Fig. 19: second row and fourth column, bottom row and last column). The contours show that the failure cases give under-segmentation, often falling inside the boundary of the tissue. This suggests that either the search path is unable to localize the boundary, or the boundary detector can be improved to more robustly classify boundaries along that search path. The use of a shallow, fully-connected boundary detector could be replaced with a Graph-CNN architecture which allows localized information from the small patches extracted along the search path to be learned.

7.4. Complexity of marginal space learning classifier

Graph convolutions, as defined in Section 5, are an element-wise multiplication of the spectral graph signal and a spectral filter, resulting in $K^{l-1}N$ trainable weights and biases per output feature map, where N is the number of vertices in the graph and K^{l-1} is the number of input feature maps. For our Graph-CNN implementation, we utilize the property that smooth spectral filters provide localized filtering in the spatial domain. Such a formulation provides the benefit of spatially localized features, and a reduction in the number of tracked weights for our network to optimize. By tracking only $n < N$ weights and interpolating the filter up to N via a smoothing kernel, we are able to reduce the number of tunable parameters in an output feature map to $K^{l-1}n$. A smaller n provides more localization, but also introduces noise in the gradient steps during back-propagation optimization (Edwards and Xie, 2016). This parametrization helps improve parameter complexity of the graph convolution for a given filter from $\mathcal{O}(n)$ to $\mathcal{O}(K)$, given a constant tracked number of weights. For NNs, full connection provides $\mathcal{O}(n)$ complexity with a separate weighting for each input feature. If utilizing standard CNNs architectures, the ability to integrate local and contextual features comes with increased complexity from a multi-branch approach (Kamnitsas et al., 2017; Kawahara and Hamarneh, 2016) with a full set of weights for each branch, or from weight sharing through dilated kernels (Wolterink et al., 2017) which learns multiple scales of the same feature. With multi-resolution patches and Graph-CNNs we are able to learn spatial features between the high and low resolution input features without tracking multiple branches for each resolution.

8. Summary

In this study we have presented a novel method for deep learning in the irregular domain of the non-uniformly sampled grid. A patch-sampling mechanic generates a single spatial domain comprised of numerous layers at differing resolutions. Through the proposed Graph-CNN operators and architecture, we are able to learn features across multiple resolutions, utilizing the intrinsic spatial relationships between features at both local and wider scales. The use of conventional CNNs in such a domain is unfeasible due to the irregular sampling used, which does not satisfy the array structured input required for regular convolutional operations. The sampling method reduces the number of input features and does not require multiple branches to a pyramid of filters or inputs, reducing the complexity of the network architecture.

In evaluating the proposed method, we present a fully automatic, deformable modeling framework for 3D aortic root segmentation in CT images. The multi-resolution sampling strategy is generalized to 3D data, forming an irregularly spaced volume sampling method. The novel segmentation pipeline method significantly reduced the time taken for training by automatically finding shape correspondence across the training set and utilizing deep learning for discriminative feature extraction, rather than hand-crafting features for the task. The MSL search space optimization benefited from a novel implementation of a multi-resolution sampling for Graph-CNN based learning of features, learning spatially related features within an irregular spatial domain. Both qualitative and quantitative results justified our proposed segmentation pipeline over a top-down approach.

Future work will explore the development of end-to-end graph-based segmentation architectures which utilize the multi-resolution approach proposed here to produce a dense segmentation of the observed domain, whether grid-based or not, making use of local and contextual information.

References

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features. Comput. Vis. Image Underst. 110, 346–359.

Boureau, Y.L., Ponce, J., LeCun, Y., 2010. A theoretical analysis of feature pooling in visual recognition. In: International Conference on Machine Learning. pp. 111–118.

Bruna, J., Zaremba, W., Szlam, A., LeCun, Y., 2014. Spectral networks and locally connected networks on graphs. In: International Conference on Learning Representations.

Caesar, H., Uijlings, J.R.R., Ferrari, V., 2016. Region-based semantic segmentation with end-to-end training. CoRR abs/1607.07671.

Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N., 2016. A unified multi-scale deep convolutional neural network for fast object detection. CoRR.

Chen, L., Yang, Y., Wang, J., Xu, W., Yuille, A.L., 2015. Attention to scale: Scale-aware semantic image segmentation. CoRR abs/1511.03339.

Cheung, A., Lichtenstein, K.M., 2012. Illustrated techniques for transapical aortic valve implantation. Ann. Cardiothorac. Surg. 1.

Cootes, T.F., Cooper, D.H., Graham, J., 1995. Active shape models - Their training and application. Comput. Vis. Image Underst. 61, 38–59.

Curcio, C.A., Sloan, K.R., Kalina, R.E., Hendrickson, A.E., 1990. Human photoreceptor topography. J. Comp. Neurol. 292, 497–523.

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In: Neural Information Processing Systems.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12, 2121–2159.

Edwards, M., Xie, X., 2016. Graph based convolutional neural networks. In: British Machine Vision Conference. pp. 114.1–114.11.

Edwards, M., Xie, X., 2017. Graph-based CNN for human action recognition from 3D pose. In: British Machine Vision Conference Workshop: Deep Learning on Irregular Domains.

Florack, L., ter Haar Romeny, B., Viergever, M., Koenderink, J., 1996. The gaussian scale-space paradigm and the multiscale local jet. Int. J. Comput. Vis. 18, 61–75.

Frangi, A.F., Rueckert, D., Schnabel, J.A., Niessen, W.J., 2002. Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling. IEEE Trans. Med. Imaging 21, 1151–1166.

Freeman, W.T., Adelson, E.H., 1991. The design and use of steerable filters. IEEE Trans. Pattern Anal. Mach. Intell. 13, 891–906.

Gao, C., Sang, N., Huang, R., 2014. Biologically inspired scene context for object detection using a single instance. PLoS One 9, 1–13.

Gessert, N., Schlüter, M., Schlaefer, A., 2018. A deep learning approach for pose estimation from volumetric OCT data. CoRR abs/1803.03852.

Ghesu, F.C., Krubasik, E., Georgescu, B., Singh, V., Zheng, Y., Hornegger, J., Comaniciu, D., 2016. Marginal space deep learning: Efficient architecture for volumetric image parsing. IEEE Trans. Med. Imaging 35, 1217–1228.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Grbić, S., Ionasec, R., Vitanovski, D., Voigt, I., Georgescu, B., Navab, N., Comaniciu, D., 2012. Complete valvular heart apparatus model from 4D cardiac CT. Med. Image Anal. 16, 1003–1014.

He, K., Zhang, X., Ren, S., Sun, J., 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European Conference on Computer Vision. pp. 346–361.

Henaff, M., Bruna, J., LeCun, Y., 2015. Deep convolutional networks on graph-structured data. CoRR abs/1506.05163 URL https://arxiv.org/abs/1506.05163.

Jiang, J., Trundle, P., Ren, J., 2010. Medical image analysis with artificial neural networks. Comput. Med. Imaging Graph. 34, 617–631.

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. Med. Image Anal. 36, 61–78.

Kawahara, J., Hamarneh, G., 2016. Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers. In: Machine Learning in Medical Imaging. pp. 164–171.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations.

Koenderink, J., 1984. The structure of images. Biol. Cybernet. 50, 363–370.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

Li, X., Yang, F., Cheng, H., Chen, J., Guo, Y., Chen, L., 2017. Ulti-scale cascade network for salient object detection. In: ACM International Conference on Multimedia. pp. 439–447.

Lienhart, R., Maydt, J., 2002. An extended set of Haar-like features for rapid object detection. In: IEEE International Conference on Image Processing. pp. 900–903. http://dx.doi.org/10.1109/ICIP.2002.1038171.

Lindeberg, T., 1996. Scale-space: A framework for handling image structures at multiple scales. In: CERN School of Computing. pp. 27–38.

Lindsay, P.H., Norman, D.A., 1977. Human Information Processing. Academic Press.

Liu, P., Wang, X., Gu, Y., 2014. Graph signal coarsening: Dimensionality reduction in irregular domain. In: IEEE Global Conference on Signal and Information Processing. pp. 798–802.

Long, J., Shelhamer, E., Darrell, T., 2014. Fully convolutional networks for semantic segmentation. CoRR abs/1411.4038.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. In: International Conference on Computer Vision. pp. 1150–1157.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. CoRR abs/1606.04797.

Monti, F., Boscaini, D., Masci, J., Rodolà, E., Svoboda, J., Bronstein, M.M., 2016. Geometric deep learning on graphs and manifolds using mixture model cnns. CoRR abs/1611.08402.

Oh, J., Kim, H.I., Park, R.H., 2017. Context-based abnormal object detection using the fully-connected conditional random fields. Pattern Recognit. Lett. 98, 16–25.

Palmer, R., Xie, X., Tam, G., 2015. Automatic aortic root segmentation with shape constraints and mesh regularisation. In: British Machine Vision Conference. pp. 83.1–83.11.

Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R., 2017. 3D Graph neural networks for RGBD semantic segmentation. In: International Conference on Computer Vision. pp. 5209–5218.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149.

Rosenholtz, R., 2016. Capabilities and limitations of peripheral vision. Ann. Rev. Vis. Sci. 2, 437–457.

Rueckert, D., Sonoda, L.I., Hayes, C., Mill, D., Leach, O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: Application to breast MR images. IEEE Trans. Med. Imaging 18, 712–721.

Safro, I., 2009. Comparison of coarsening schemes for multilevel graph partitioning. In: International Conference on Learning and Intelligent Optimization. pp. 191–205.

Safro, I., Sanders, P., Schulz, C., 2012. Proceedings of the Interntional Symposium on Experimental Algorithms. Springer, pp. 369–380, chapter Advanced Coarsening Schemes for Graph Partitioning.

Shin, H.C., Orton, M.R., Collins, D.J., Doran, S.J., Leach, M.O., 2013. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1930–1943.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35, 1285–1298.

Shuman, D.I., Faraji, M.J., Vandergheynst, P., 2016. A multiscale pyramid transform for graph signals. IEEE Trans. Signal Process. 64, 2119–2134.

Shuman, D., Narang, S., Frossard, P., Ortega, A., Vandergheynst, P., 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process. Mag. 30, 83–98.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Tony, L., 2008. Wiley Encyclopedia of Computer Science and Engineering. John Wiley and Sons, pp. 2495–2504, chapter Scale-Space.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 511–518.

Wang, P., Gan, Y., Shui, P., Yu, F., Zhang, Y., Chen, S., Sun, Z., 2018. 3D Shape segmentation via shape fully convolutional networks. Comput. Graph. 70, 128–139.

Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I., 2017. Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease. In: Joint International Workshop on Reconstruction and Analysis of Moving Body Organs, and Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease. pp. 95–102.

Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X., 2017. SegAN: Adversarial network with multi-scale L_1 loss for medical image segmentation. CoRR abs/1706.01805.

Yan, S., Smith, J.S., Lu, W., Zhang, B., 2017. Hierarchical multi-scale attention networks for action recognition. CoRR abs/1708.07590.

Ying, R., You, J., Morris, C., Ren, X., Hamilton, W.L., Leskovec, J., 2018. Hierarchical graph representation learning with differentiable pooling. CoRR abs/1806.08804.

Yosinski, J., Clune, J., Nguyen, A.M., Fuchs, T., Lipson, H., 2015. Understanding neural networks through deep visualization. CoRR abs/1506.06579.

Yu, R., Chen, X., Morariu, V.I., Davis, L.S., 2016. The role of context selection in object detection. CoRR abs/1609.02948.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. pp. 818–833, European Conference on Computer Vision.

Zhang, J., Dai, Y., Li, B., He, M., 2017. Attention to the scale: Deep multi-scale salient object detection. In: International Conference on Digital Image Computing: Techniques and Applications. pp. 1–7.

Zhang, Z.V., Tang, M., Cobzas, D., Zonoobi, D., Jägersand, J.L., 2018. End-to-end detection-segmentation network with ROI convolution. CoRR abs/1801.02722.

Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D., 2008. Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features. IEEE Trans. Med. Imaging 27, 1668–1681.

Zheng, Y., Comaniciu, D., 2014. Marginal Space Learning for Medical Image Analysis. Springer.

Zheng, Y., Georgescu, B., Ling, H., Zhou, S.K., Scheuering, M., Comaniciu, D., 2009. Constrained marginal space learning for efficient 3D anatomical structure detection in medical images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 194–201.

Zheng, Y., John, M., Liao, R., Nottling, A., Boese, J., Kempfert, J., Walther, T., Brockmann, G., Comaniciu, D., 2012. Automatic aorta segmentation and valve landmark detection in c-arm CT for transcatheter aortic valve implantation. IEEE Trans. Med. Imaging 31, 2307–2321.