sensors

MDPI

Article

Myocardium Detection by Deep SSAE Feature and Within-Class Neighborhood Preserved Support Vector Classifier and Regressor

Yanmin Niu [1,2,*], Lan Qin [1] and Xuchu Wang [1]

[1] Key Laboratory of Optoelectronic Technology and Systems of Ministry of Education, College of Optoelectronic Engineering, Chongqing University, Chongqing 400044, China; qinlan@cqu.edu.cn (L.Q.); xcwang@cqu.edu.cn (X.W.)

[2] College of Computer and Information Science, Chongqing Normal University, Chongqing 400050, China

* Correspondence: niuym@cqnu.edu.cn

Received: 1 March 2019; Accepted: 28 March 2019; Published: 13 April 2019

Abstract: Automatic detection of left ventricle myocardium is essential to subsequent cardiac image registration and tissue segmentation. However, it is considered challenging mainly because of the complex and varying shape of the myocardium and surrounding tissues across slices and phases. In this study, a hybrid model is proposed to detect myocardium in cardiac magnetic resonance (MR) images combining region proposal and deep feature classification and regression. The model firstly generates candidate regions using new structural similarity-enhanced supervoxel over-segmentation plus hierarchical clustering. Then it adopts a deep stacked sparse autoencoder (SSAE) network to learn the discriminative deep feature to represent the regions. Finally, the features are fed to train a novel nonlinear within-class neighborhood preserved soft margin support vector (*C*-SVC) classifier and multiple-output support vector (ε-SVR) regressor for refining the location of myocardium. To improve the stability and generalization, the model also takes hard negative sample mining strategy to fine-tune the SSAE and the classifier. The proposed model with impacts of different components were extensively evaluated and compared to related methods on public cardiac data set. Experimental results verified the effectiveness of proposed integrated components, and demonstrated that it was robust in myocardium localization and outperformed the state-of-the-art methods in terms of typical metrics. This study would be beneficial in some cardiac image processing such as region-of-interest cropping and left ventricle volume measurement.

Keywords: myocardium detection; cardiac magnetic resonance; region proposal; support vector classifier and regressor; stacked sparse autoencoder (SSAE)

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death and disability globally. For years, a great effort has been dedicated to the prevention, diagnosis, treatment and research of CVDs. The hardware and software developments have been helping the increasing use of cardiovascular magnetic resonance imaging (MRI) in this effort. It is essential to detect the important structures of a left ventricle myocardium from MRI scans in a clinical-decision support system dedicated to improving the early diagnosis of critical CVD diseases. For example, accurate myocardium location will be very helpful for subsequent processing such as cardiac image registration and tissue segmentation, also for understanding cardiac anatomy how to adapts to disease [1]. Computer-aided automatic detection provides great potential to solve this problem instead of tedious, time-consuming, and poorly reproducible manual detection. However, this has been a challenging task due to the complex structure of cardiac anatomy, and low image quality such as presence of noise, low contrast and intensity non-uniformity [2–4].

1.1. Related Works

Myocardium detection is a task that has benefited from the object detection in the computer vision field. Traditionally, hand crafted features, such as HOG (histogram of oriented gradients), SIFT (scale-invariant feature transform), Haar-like feature, etc, are widely used to train various classifiers [5]. The enhanced cascade detector [6] that was originally developed for face detection, and the decision forest detector that combines a wide range of contextual characteristics and random forest classifier to locate nine different organizations on human body images [7].

However, along with recent breakthrough works in deep learning field, many CNN (Convolutional Neural Network) architectures have been studied for object detection [8–12] and achieved more satisfactory performance in nature or optical images. Typically, these works can be roughly divided into two categories: region proposal based methods and region proposal free methods. The former mainly includes RCNN (Regional Convolutional Neural Network) [8], SPP-Net (Spatial Pyramid Pooling Net) [13], Fast-RCNN [14], Faster-RCNN [9], R-FCN (Region-based Fully Convolutional Network) [15] and its multi-scale version [16] and cascaded improvement [17]. Furthermore, the relation among the detected objects is modeled by a CNN network with two full connected layers [18], and the iterative localization refinement is designed to facilitate object localization by undertaking at a mid-layer of a CNN to progressively refines a subset of region proposals[19]. The second category is not using the region proposal, such as three versions of YOLO (You Only Look Once) [10,20,21] and SSD (Single Shot MultiBox Detector) [22] and its improvement [23].

In this work, we focus on region proposal-based methods since they are applied by most of the top-performing object detection methods. In this two-stage method, a sparse set of candidate regions is first generated, and then they are further classified and regressed. The representative RCNN built the relationship between image classification and object detection by three steps: First, selective search [24] is applied to generate around 2000 category-independent region proposals in stead of the traditional sliding window approach. Second, the features of each region proposal are extracted by a pretrained CNN model. Third, the top-level features are classified by linear SVM (Support Vector Machines). RCNN has a solid pipeline but its computation speed is slow because it performs a CNN forward pass for each object proposal, without sharing computation. Fast RCNN [14] combines the region proposal classification and bounding box regression tasks into one single stage to speed up the detection. Moreover, the region of interest pooling strategy based on the top-level features is more efficient than the RCNN feature extracting method. In other words, multi-task training avoids managing a pipeline of sequentially-trained tasks. Nevertheless, because selective search is applied to generate region proposals in Fast RCNN, thereafter the detection speed of Fast RCNN is affected. Faster RCNN [9] solves the proposal computation bottleneck of Fast RCNN by using a region proposal network that is a kind of fully convolutional network and can be trained end-to-end to generate detection proposals.

Generally, these works mainly aim to improve the object detection accuracies in two ways: (1) optimizing the architecture of the CNN, take full advantages of the distinguished ability of the deep feature learned by CNN; (2) exploring how to share computation among different proposed regions, which will speed up the whole detection process. Along with these strategies, some deep learning techniques have been applied for medical object detection, for example, Yan et al. [25] uses a system containing two convolutions depth convolution neural network, with 7000 two-dimensional axis of the slice image training, and ultimately to the body of 12 different organizations. Vos et al. [26] trained three independent CNNs based ROI detectors, where each classified 2D image slices from one of three orthogonal image planes (axial, sagittal, or coronal), then all of them were combined to determine a rectangular 3D bounding of anatomical ROI. This method achieved good detection results, but because of the need to train multiple networks, the algorithm itself was less efficient. Roth et al. [27] trained a 5-layer convolution neural network by using 4300 two-dimensional axial-shaped slices on the human body of five different regions (legs, pelvic, liver, lung, and neck). In the application to cardiac tissue detection, Luo et al. [28] employed a 8-layer fully convolutional networks to locate the

ROI (Region Of Interest) that contains the bi-ventricular regions for right ventricle segmentation. Poudel et al. [29] proposed a recurrent fully-convolutional network that combines left ventricle detection and segmentation into a single architecture that is trained end-to-end thus simplifying the segmentation pipeline. Tan et al. [30] proposed a CNN network with fully connected layer to regress the location of left ventricle center point in cardiac images. Vigneault et al. [4] built a small localization network from the layer immediately following the final max pooling of U-Net to predict the transformation parameters in locating the left ventricle in cardiac images. This method can output the directed localization but contains much more surrounding tissues. Overall speaking, the investigation on myocardium detection is very limited in comparison to the advances in natural object detection and most existing approaches usually take it as a module in the pipeline of cardiac segmentation, where they strengthen the completeness of myocardium more remarkably than the accuracy.

1.2. Motivation and Contribution

In above methods, Faster RCNN obtains both high detection efficiency and detection accuracy without changing the pipeline of region proposal and region classification. Unfortunately, four problems are not solved in the studies. Firstly, the network training requires huge amount of labeled training data, which makes it hard for medical image applications to utilize this technology, due to the fact that it is extremely hard to collect such a large data with correctly diagnosed labels. Secondly, contextual information is not integrated with the top-level features. Thus, the quality of generated region proposals is relatively poor. Thirdly, the design for the selected scales and aspect ratios of anchor boxes is not optimal for medical objects because in our task there is some anatomical constraints that cannot be described as the limited scales and aspect ratios. Therefore, the ability of regional proposal network object localization is weak for myocardium detection. Fourthly, the classifier is not optimal for solving binary classification problem since it does not consider the structure information of the top-level features. As a result, the performance is affected in detection of specific medical tissues.

The objective we consider is to localize myocardium and left ventricle tissue (or, left ventricle ROI) in cardiac MRI images, where the object is single and the anatomical information is more remarkable in comparison to multiple objects in natural images. However, the localization remains a challenging problem due both to intrinsic and extrinsic difficulties. Intrinsic difficulties refer to the essential properties of the MR imaging systems that result in imaging noise and the complexity of cardiac tissues. The major source of noise that degrades image quality is mainly caused by radiation scattering and source leakage. The complexity of cardiac tissues lies on that in typical short-axis steady-state free precession cine MR images, the contrast between the blood pool within the left ventricle and the endocardial wall is varying, the interference of endocardial trabeculation and papillary muscles is relative strong, and the contrast between the epicardial wall and surrounding structures is extremely weak and varying, particularly against low-signal lung tissue. Extrinsic difficulties are closely related to the patients with biological variability in heart size, orientation in the thorax, and morphology across subjects. Also there is variability in contrast and image appearance with different scanners, protocols, and clinical planes. In some images, the borders between the ventricles and the atria, and the separation between the chambers and the vasculature are hard to define. In this context, it is admirable to generate candidate regions according to the texture and structure information of the cardiac image, instead of the convolution network-based region proposal network. Also a stronger classifier is necessarily investigated to distinguish the proper proposal regions from large mount of candidates.

Motivated by these observations, in this paper, we propose a hybrid model to detect myocardium by integrating region proposal and deep feature classification and regression. Our model firstly generates candidate regions on cardiac images using new structural similarity-enhanced supervoxel over-segmentation plus hierarchical clustering. Compared with the typical sliding window extraction methods, the algorithm is more efficient and generates less redundant regions. Then it adopts a deep SSAE (Stacked Sparse Auto-Encoder) network to learn the deep feature to represent the candidate

regions. Compared with the traditional manual feature, the difference between the myocardium region and background is strengthened by the supervised SSAE deep learning, which also improves the robustness of our model. Finally, the obtained deep feature represented candidate regions are fed to train a nonlinear within-class neighborhood preserved soft margin support vector classifier (C-SVC) and multiple output support vector regressor (ε-SVR), and finally output the myocardium region. To improve the stability and generalization, the proposed model also takes the hard negative sample mining strategy to fine-tune the SSAE and the classifier. Experimental results show that the accurate myocardium detection results can be achieved by fine-tuning a pre-trained deep learning network.

Briefly speaking, the contributions and advantages of our method are highlighted as follows:

(1) The region proposal is generated using new structural similarity-enhanced supervoxel over-segmentation plus hierarchical clustering. Instead of the color and spatial similarity computation in the widely adopted SLIC (Simple Linear Iterative Clustering) method, the proposed supervoxel introduced a sequence of measure, including image phase congruency, intensity, contrast, structure, and coordinates to enhance the structural similarity, which can easily take into account the context of similar anatomical tissues while limiting the capacity of redundant proposals.

(2) The SSAE network is introduced to learn the deep features related to region proposals. SSAE enjoys all the benefits of any deep network of greater expressive power by capturing a useful "hierarchical grouping" or "part-whole decomposition" of the candidate regions. For supervised SSAE, the gradients from the softmax classification error will then be back-propagated into the encoding layers and which can enhance the difference of feature representation for positive and negative regions. Furthermore, SSAE outputs less feature by dimensionality reduction to help training more robust classifier and regressor in the successive steps.

(3) The nonlinear within-class neighborhood preserved C-SVC classifier and ε-SVR regressor are proposed to classify the SSAE-learned features and regress the locations of features to refined positions, respectively. Since there are common components in many features due to their intensity representation of limited anatomical types of cardiac images, these components will be encoded as more similar parts after SSAE learning. According to this, we propose a nonlinear regularization to preserve the within-class neighborhood structure and incorporate it to C-SVM and ε-SVR. It can alleviate the limitations that standard C-SVM and ε-SVR suffer from the noisy data that heavily affect the hyperplane, since they obtain larger non-zero coefficients after training. In addition, our multiple-input multiple-output ε-SVR can produce more robust nonlinear regression than the linear regression in many region-based detectors.

We intensively investigated the performance of the proposed method with impacts of different components and compared it with related methods on public available cardiac data set. Although Faster RCNN has become one of the most outstanding detection methods for natural images, we show that the proposed method can achieve competitive results more efficiently and has potential as well when incorporating the integration of enhanced supervoxel-based region proposal, deep learned SSAE feature, and nonlinear within-class neighborhood preserved C-SVC classifier and ε-SVR regressor. In this context, the proposed model is valuable as a reference for segmenting other similar medical objects in limited image sets.

The rest of the paper is organized as follows. Section 2 describes the flowchart of proposed model and the details of major modules. Section 3 presents the data set and our evaluation metrics. Experimental results and discussion are reported in Section 4 and we summarize our work in Section 5.

2. Proposed Method

2.1. Overview of Our Detection Model

We formulate the myocardium detection as a classification and regression problem and the training flowchart of our model is shown in Figure 1. This model consists of four modules: (1) candidate region extraction module that combines structural similarity-enhanced supervoxel algorithm and

hierarchical clustering to generate target candidate regions; (2) feature learning module that extracts the deep SSAE characteristics of the candidate regions; (3) region location module that trains a within-class neighborhood preserved *C*-SVC classifier to determine and locate the myocardium region; (4) refinement module that collects hard-to-be-classified samples and use them to fine tune the model, also eliminates redundant (cross-repeated) candidate regions and find the best target detection position by using the NMS (Non-Maximum Suppression) and ε-SVR bounding box regression strategies. In the following subsections, we will present the details of each module.

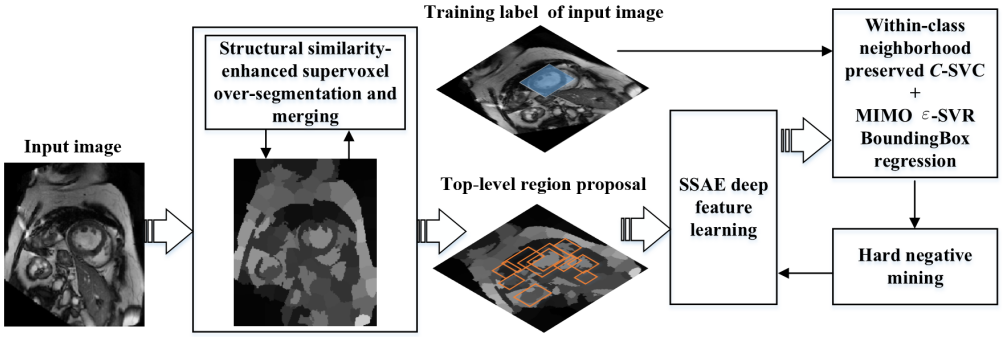

Figure 1. Training flowchart of proposed myocardium detection model.

2.2. Candidate Region Proposal

In two-stage object detection methods, it is usual to generate the candidates in possible sizes, scales, locations to handle the strong randomness of target located on the image. The first early region candidate generation algorithm adopted sliding window [6] to traverse the entire images, the image needs to be set as different scales and sliding window aspect ratio. This exhaustive search certainly could find the target, meanwhile it generated more redundant proposals with the much high time complexity, furthermore, the huge negative proposals makes the classifier less sensitive to the positive proposals. To overcome the limitation, selective search [24], edge boxes [31], region proposal network [9], superpixel proposal [32–34] were introduced to generate candidate regions quickly and efficiently. For our detection task, there are many similar anatomical structures in the limited cardiac images, so we consider the supervoxel-based over-segmentation to generate the initial regions.

2.2.1. Structural Similarity-Enhanced Supervoxel Over-Segmentation

The distance or similarity measure plays an essential role in supervoxel framwork. For typical SLIC, the feature for distance measure is built as LAB-color space-based intensities balanced with the pixel location distance [33,35]. No structural information is incorporated. For cardiac image, this may be less appropriate since the intensity is single channel and objects are usually with coarse boundaries. So we introduce five parts to enhance the structural similarity as follows.

(1) Phase congruency measure

Image phase congruency (PhaseCong) model assumes the visual feature should be high in information (or entropy), and low in redundancy. Instead of searching for points where there are sharp changes in intensity, this model searches for patterns of order in the phase component of the Fourier transform. Based on the physiological and psychophysical evidences, the PhaseCong theory provides a simple but biologically plausible model of how mammalian visual systems detect and identify features in an image. Rather than define features directly at points with sharp changes in intensity, the PhaseCong model postulates that features are perceived at points where the Fourier components are maximal in phase. PhaseCong can be considered as a dimensionless measure for the significance of a local structure [36].

To compute the PhaseCong of cardiac images, we can apply the two-dimensional log-Gabor filters that uses the Gaussian spreading function across the filter perpendicular to its orientation. In this way,

the phase of any function would stay unaffected after being smoothed with Gaussian. Thus, the phase congruency would be preserved. This function has the following transfer formulation

$$G(\omega,\theta) = \exp\left(-\frac{log(\omega/\omega_0)^2}{2\sigma_r^2}\right)\exp\left(-\frac{(\theta-\theta_j)^2}{2\sigma_\theta^2}\right), \tag{1}$$

where ω_0 is the filter's center frequency and σ_r controls the filter's bandwidth; θ_j is the orientation angle of the angle filter and σ_θ determines the filter's angular bandwidth. By modulating ω_0 and θ_j and convolving G with the image, a set of responses at each point u as $[e_{n,\theta_j}(u), o_{n,\theta_j}(u)]$. The local amplitude on scale n and orientation θ_j is $A_{n,\theta_j}(u) = \sqrt{(e_{n,\theta_j}(u))^2 + (o_{n,\theta_j}(u))^2}$ and the local energy along orientation θ_j is $E_{n,\theta_j}(u) = \sqrt{(\sum_n e_{n,\theta_j}(u))^2 + (\sum_n o_{n,\theta_j}(u))^2}$, therefore, the phase congruency at the point u is obtained as

$$PhaseCong(u) = \frac{\sum_j E_{\theta_j}(u)}{\sum_n \sum_j A_{n,\theta_j}(u)}, \tag{2}$$

and the phase congruency measure of two points are defined as

$$S_{pm}(u,v) = \frac{2PhaseCong(u)PhaseCong(v) + c_1}{PhaseCong(u)^2 + PhaseCong(v)^2 + c_1}. \tag{3}$$

(2) Intensity measure

For two patches $Pat(u)$ and $Pat(v)$ centered at position u and v, with $d = p \times p$ patch size in cardiac image ($p = 7$ or 9 is better in our study), we define the intensity measure as $S_{mm}(u,v) = \frac{2\mu_u\mu_v + c_2}{\mu_u^2 + \mu_v^2 + c_2}$, where $\mu_u = 1/d\sum_{i=1}^{d} Pat_i(u)$ and $\mu_v = 1/d\sum_{i=1}^{d} Pat_i(v)$ are the mean intensities of the compared patches, and constant c_2 is included to avoid instability when the intensities of two patches are near to zero. If the mean intensities of two patches are close, $S_{mm}(u,v)$ will approach to 1 and vice versa.

(3) Contrast measure

Once the mean intensity is removed from each patch, the resulting signal can be seen as the inner contrast of the patches, so we use the standard deviation to estimate the similarity of these contrast, i.e., $S_{cm}(u,v) = \frac{2\sigma_u\sigma_v + c_3}{\sigma_u^2 + \sigma_v^2 + c_3}$, where $\sigma_u^2 = 1/(d-1)\sum_{i=1}^{d}(Pat_i(u) - \mu_x)^2$ and $\sigma_v^2 = 1/(d-1)\sum_{i=1}^{d}(Pat_i(v) - \mu_v)^2$, c_3 plays the same role as c_2. If the mean contrast of two patches are close, $S_{cm}(u,v)$ will approach to 1 and vice versa. Furthermore, this measure is less sensitive to the case of high base contrast than low base contrast and consistent with the contrast-masking feature of the human visual system.

(4) Structure measure

For two patches $Pat(u)$, $Pat(v)$ centered at u and v, correlation (inner product) between them is a simple but effective measure to quantify their structural similarity. Since it equals to the correlation coefficient of the normalized patches with voxels $(Pat_i(u) - \mu_u)/\sigma_u$ and $(Pat_i(v) - \mu_v)/\sigma_v$, so we define the structure measure as $S_{sm}(u,v) = \frac{2\sigma_{uv} + c_4}{\sigma_u\sigma_v + c_4}$, where $\sigma_{uv} = 1/(n-1)\sum_{i=1}^{d}(Pat_i(u) - \mu_u)(Pat_i(v) - \mu_v)$. Geometrically, the correlation coefficient corresponds to the cosine of the angle between the vectors with elements $(Pat_i(u) - \mu_u)$ and $(Pat_i(v) - \mu_v)$, so we take the absolute operation to constrain it into the range of 0 and 1.

(5) Coordinate measure

The coordinate measure is defined as $S_{dm}(u,v) = \exp(-\alpha||u-v||^2)$, where α is related to the supervoxel number K and image voxel number N. Typically, $\alpha = 2\sqrt{K/N}$. Different from the distance measure in SLIC method, this definition explicitly constrains the similarity into $[0,1]$.

Then, these measurements are combined to get the hybrid similarity of patches centered at the position u and v on the image. We define $S(u,v)$ as

$$\begin{aligned} S(u,v) =& [S_{pm}(u,v)]^{\beta_1}[S_{im}(u,v)]^{\beta_2} \\ & [S_{cm}(u,v)]^{\beta_3}[S_{sm}(u,v)]^{\beta_4}[S_{dm}(u,v)]^{\beta_5}, \end{aligned} \tag{4}$$

where $\beta_i (i = 1, \cdots, 5)$ are weights for the corresponding part, in our experiments, we let them as 1 for simplicity. This comprehensive distance measure explicitly integrates the different measures of two patches into a unified measure space, and it can be separately calculated and then incorporated to form the final results for speeding up the computation.

2.2.2. Supervoxel Region Merging by Hierarchical Clustering

The initial over-segmentation regions by the supervoxel algorithm usually divide the objects into many adjoint parts. These regions should be adjusted to represent the object more accurately and efficiently. We introduce hierarchical clustering algorithm to merge the generated regions in a bottom-up way. Each of the two merged regions satisfies two conditions: (1) the regions should be adjacent; (2) the two regions have the highest similarity. The similarity is computed using Equation (4), but in a supervoxel region instead of patch, and the coordinate measure are computed in the center of the supervoxel instead of locations of voxels. Theoretically, the final result of clustering is that all the initial regions are merged into the same region (that is, the whole image). So, it is necessary to set the desired number of final regions, so that a candidate target area for detection can be finally obtained.

The complete candidate region generation algorithm is described as follows:

Step 1: Generating initial over-segmentation regions by supervoxels framework with the similarity measure in Equation (4) in a patch-wise way;

Step 2: Calculating similarity between all adjacent regions by using the similarity measure in Equation (4) in a supervoxel-wise way;

Step 3: Sorting the similarities, and then merging the two regions with the highest similarity to form new over-segmentation regions;

Step 4: Repeat Steps 2 and 3 until the number of remaining areas reaches to the predefined value.

Figure 2 illustrates the procedure of our region proposal generation on three typical images located in the base, middle, and apex parts of myocardium tissue, respectively. It is seen that the targets both include in-homogeneous blood pools and myocardium and show much variability in complex background. The endocardium is not always closed circle while the epicardium exhibits obscure boundaries to surrounding tissues. Our initial over-segmentation extract the anatomical structures in most cases, but still separate the targets into different adjacent parts. After hierarchical clustering this disjoint limitation is greatly decreased, and the final region proposals covers the true objects through some restrictions, e.g., the ratio of height and width in bounding box, the number of voxels in supervoxels, the location near to the image border. It is also noticeable that our approach generates much less proposal than selective search-based methods.

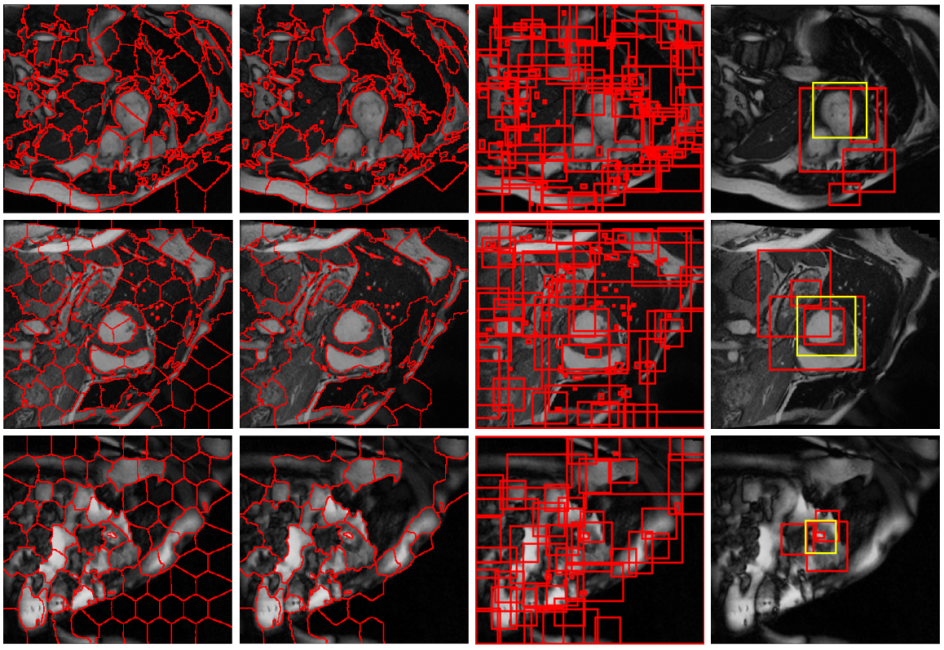

Figure 2. Procedure of our region proposal generation. From left to right: initial structural similarity-enhanced supervoxels; merged supervoxels by hierarchical clustering; corresponding bounding boxes; final top-level region proposals where yellow rectangles denote ground truth.

2.3. Deep SSAE Feature Learning

It is difficult to design a hand-crafted feature to capture the characteristics of left ventricle and myocardium tissue in different cardiac images because the gray scales of left ventricle target in heart MRI images are varying in complex morphological changes and image background. For learning-based detection model, an effective feature representation can relieve this burden. In view of the excellent ability and robustness of the deep learning characteristics, this paper propose to extract deep feature representation of the proposal regions using deep SSAE model.

To be specific, the region proposal generation algorithm outputs candidates with different scales and different sizes according to supervoxel and regional hierarchical clustering. For each candidate, the minimum bounding rectangle is built and the corresponding region from the original image is cut down to form a training sample. Since the dimensions of bounding rectangles are different across the images, they should be scaled to a fixed size (i.e., $\tau \times \tau$). For SSAE, the number of neurons in the input layer is the dimension of the training sample, so the value of τ^2 is directly determined by the number of neurons in the SSAE input layer, which is one of the important parameters of SSAE structure. In our experiments, the value of τ is settled via cross-validation.

SSAE is a deep neural network composed of multiple stacked sparse auto-encoders (SAEs) [37], and it has been applied in tissue segmentation in late gadolinium-enhanced cardiac MRI images (such as atrial scarring segmentation [38], atrial fibrosis segmentation [39], left atrium segmentation [40]), nuclei patch classification on breast cancer histopathology images [41], brain tissue segmentation in visible human images [42], or other applications (such as hyperspectral imagery classification [43] and building extraction from LiDAR and optical images [44]). Figure 3 shows a SSAE network with three hidden layers, where a SAE aims to learn features that form a good sparse representation of its input. The first layer of a SSAE tends to learn first-order features in the raw input (such as edges in a proposal). The second layer tends to learn second-order features corresponding to patterns in the appearance of first-order features (e.g., in terms of what edges tend to occur together). Higher layers of the SSAE tend to learn the sparse but even higher-order features, which can be admirable for classifying myocardium regions.

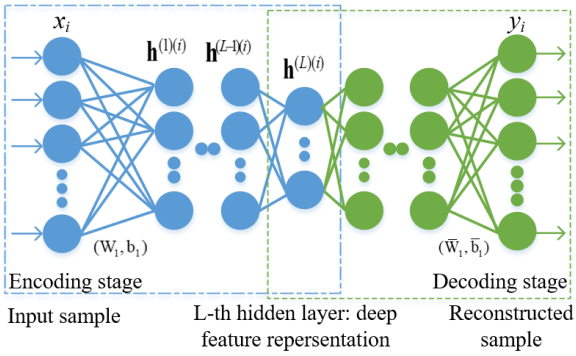

Figure 3. SSAE structure with three hidden layers in this work.

SSAE includes encoding and decoding phases and each phase contains more layers (here, three, which can be set according to specific tasks). Given an input sample x, the first SAE maps it to the activation vector $h^{(1)(i)}(x_i) = f(W_1 x_i + b_1)$, where $f(z) = 1/(1 + \exp(-z))$ is the sigmoid function to make non-linear activation, $W_1 \in R^{H \times M}$ represents the coding weight matrix in the first layer of the sparse self-encoder, b_1 represents the offset variable; then, this vector $h^{(1)(i)}(x_i)$ is used as the input vector for the second SAE mapped to activation vector $h^{(2)(i)}(x_i)$; in a similar way, this vector is finally expressed as the final depth characteristic of the input sample. The average hidden layer activation value of this neuron for all training samples can be expressed as $\hat{\rho}_j = \frac{1}{n}\sum_{i=1}^{n} h_j(x^{(i)})$, in this condition, we can set a small sparse parameter (e.g., $\rho = 0.01$), by making $\hat{\rho}_j = \rho$, the mean activation of each neuron will be close to ρ, so as to achieve the purpose of sparse constraints and it can be formulated as

$$\sum_{j=1}^{s_2} KL(\rho||\hat{\rho}_j) = \sum_{j=1}^{s_2} \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}. \tag{5}$$

Here, s_2 represents the number of neurons in the implicit layer, and the index j represents the jth hidden layer neurons. $KL(\rho||\hat{\rho}_j)$ is a Kullback-Leibler divergence that measures the similarity between two different distributions ρ and $\hat{\rho}_j$.

The overall cost function of SSAE is defined as

$$J = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}||y^{(i)} - x^{(i)}||^2\right) + \lambda \frac{1}{2}\sum_{l=1}^{p-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l-1}}\left(W_{ji}^{(l)}\right)^2 + \beta\sum_{j=1}^{s_2} KL(\rho||\hat{\rho}_j), \tag{6}$$

where the first term is defined as a mean square error cost function to learn an identity function so that output y_i equals to input x_i. The second one is a quadratic regularized function to penalize the parameters W and reduce the risk of the model being over-fitted. The third term is a sparse constraint.

The training of SSAE is the process of optimizing its cost function. It is generally conducted using step-by-step greedy strategy. It consists of two parts: (1) model pre-training (progressive training for each SAE); (2) model fine-tuning (fine-tuning of the pre-trained model). In the first stage, the training samples are used to train the first layer of SAE separately. After this training is completed, the second layer SAE is trained by using the hidden layer output activation vector of the first layer SAE, and then the output activation vector of the second layer SAE is used to the training of the third layer of SAE (if there are more hidden layers, the training processes in the same way). After the first part of the training is completed, a number of trained SAE is stacked into a multi-layer SSAE, the initial parameters of SSAE are feed by the weight parameters of SAE, and the last hidden layer of neurons connects to the classifier to form a complete classification network. The gradient descent algorithm is used to pass the error of the Softmax classifier to the entire network, and the whole network is then fine-tuned [37,45].

2.4. Within-Class Neighborhood Preserved C-SVC Classification

In this paper, the two-class soft margin support vector classifier is considered to classify the candidate regions (each candidate region is divided into a target area or a non-target area). Standard *C*-SVM suffers from the noisy data that heavily affect the hyperplane, since they obtain larger non-zero coefficients after training [46]. In addition, the feature learned by SSAE is biased for the negative samples, that are the mixture of the complex background. To alleviate these limitations, we propose to incorporate local geometric structure to constrain the maximum margin-based *C*-SVC, and the classifier is built as follows.

Suppose there are N SSAE learned samples $\mathbf{x}_i (i = 1, \cdots, N)$ and they belong to class $\mathcal{C}_k (k = 1, 2)$, and the size of each class is N_k. There is a linear or nonlinear mapping ϕ to transform $\mathbf{x}$ into an arbitrary reproducing kernel Hilbert space $\mathcal{H}$, i.e., $\phi : \mathbb{R}^D \mapsto \mathcal{H}$, then, according to the Mercer theorem, a kernel function $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ could be designed to avoid curse of dimensionality. Exploiting the manifold in the form of a graph can be seen as a method of incorporating local proximity information of the images into the dimensionality reduction framework, that can enhance the clustering quality in the low-dimensional space. Assume the mapped data are centered in $\mathcal{H}$, i.e., $\sum_{i=1}^{N} \phi(\mathbf{x}_i) = 0$, and total scatter matrix of samples in feature space is $\mathbf{S}_t^{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^T$, the eigen-decomposition is $\mathbf{S}_t^{\phi} \mathbf{v} = \lambda \mathbf{v}$; $\mathbf{v} \in span\{\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_N)\}$. Let $\mathbf{v} = \sum_{i=1}^{N} \mu_i \phi(\mathbf{x}_i)$, we obtain $\lambda N \mu = \mathbf{K} \mu$ due to $\mathbf{K}$ is symmetric and has a set of eigenvectors spanning the whole space. In case that the mapped data are not centered in $\mathcal{H}$, we replace $\mathbf{K}$ by $(\mathbf{I} - \mathbf{e}\mathbf{e}^T)\mathbf{K}(\mathbf{I} - \mathbf{e}\mathbf{e}^T)$ to implement centralization, where $\mathbf{e} = N^{-1/2} \mathbf{1}_N^T$. In this way, a sample $\phi(\mathbf{x}_i)$ obtains its projection as

$$\phi(\mathbf{x}_i)_{kpca} = \mathbf{W}_{kpca}^T \mathbf{K}(\mathbf{x}_i, \cdot), \tag{7}$$

where $\mathbf{W}_{kpca} = [\mu_1, \mu_2, \cdots, \mu_{N-1}]$ with arrangement of μ_i according to the descending order of eigenvalues. This KPCA preprocessing does not lose any information due to the representation theorem and the orthogonal decomposition technique [47].

Let $\mathcal{G}$ be a graph built on samples $\mathbf{X}^{\phi} = [\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_N)]$ and $\mathbf{A}$ be a symmetric matrix that encodes the weighted adjacency information among images, that is,

$$\mathrm{A}_{ij} = \begin{cases} \frac{1}{\mathrm{D}_{i.}} \exp\left(-\frac{||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2}{\sigma_i^{(t)} \sigma_j^{(t)}} \right), & \phi(\mathbf{x}_i) \in \mathcal{N}_K(\phi(\mathbf{x}_j)) \text{ or} \\ & \phi(\mathbf{x}_j) \in \mathcal{N}_K(\mathbf{x}_i); \\ 0, & \text{other}, \end{cases} \tag{8}$$

where $\mathrm{D}_{i.} = \sum_j \mathrm{A}_{ij}$ normalizes each weight and $\sigma_i^{(t)}$ (or $\sigma_j^{(t)}$) is the distance of $\phi(\mathbf{x}_i)$ (or $\phi(\mathbf{x}_j)$) and its t-th within-class neighbor, in our experiments, $t = 7$. This settlement is more controllable than the traditional selection (e.g., the variance or predefined fixed value). By introducing the kernel function, this weight is rewritten as

$$\mathrm{A}_{ij}^{\phi} = \frac{1}{\mathrm{D}_{i.}^{\phi}} \exp\left(-\frac{\mathrm{K}_{ii} + \mathrm{K}_{jj} - 2\mathrm{K}_{ij}}{\sqrt{(\mathrm{K}_{ii} + \mathrm{K}_{i^{(t)}i^{(t)}} - 2\mathrm{K}_{ii^{(t)}})(\mathrm{K}_{jj} + \mathrm{K}_{j^{(t)}j^{(t)}} - 2\mathrm{K}_{jj^{(t)}})}} \right), \tag{9}$$

if $\phi(\mathbf{x}_i) \in \mathcal{N}_K(\phi(\mathbf{x}_j))$ or $\phi(\mathbf{x}_j) \in \mathcal{N}_K(\phi(\mathbf{x}_i))$ and $A_{i,j}^{\phi} = 0$ otherwise, where $i^{(t)}$ (or $j^{(t)}$) denotes the subscript of the t-th within-class neighbor of $\phi(\mathbf{x}_i)$ (or $\phi(\mathbf{x}_j)$). $\mathrm{D}_{i.}^{\phi} = \sum_j \mathrm{A}_{ij}^{\phi}$ is a normalizer. Based on this, we want to model the local intrinsic geometry structures and define a within-class neighborhood preserving scatter matrix in KPCA feature space as

$$\begin{aligned}\mathbf{S}_w^{\phi} &= \sum_{k=1}^{2}\sum_{\mathbf{x}_i\in\mathcal{C}_k}(\phi(\mathbf{x}_i)_{kpca} - \sum_{j=1}^{N_k}\mathrm{A}_{ij}^{\phi}\phi(\mathbf{x}_j)_{kpca})(\phi(\mathbf{x}_i)_{kpca} - \sum_{j=1}^{N_k}\mathrm{A}_{ij}^{\phi}\phi(\mathbf{x}_j)_{kpca})^T \\ &= \sum_{k=1}^{2}\mathbf{W}_{kpca}^T\mathbf{K}_k(\mathbf{I}_k - \mathbf{A}_k^{\phi})^T(\mathbf{I}_k - \mathbf{A}_k^{\phi})(\mathbf{K}_k)^T\mathbf{W}_{kpca} \\ &= \mathbf{W}_{kpca}^T\mathbf{K}(\mathbf{I} - \mathbf{A}^{\phi})^T(\mathbf{I} - \mathbf{A}^{\phi})\mathbf{K}^T\mathbf{W}_{kpca}\end{aligned}$$

where $\mathbf{K}_k$ is $N \times N$ kernel matrix whose first N_k columns and first N_k rows based block are taken from $\mathbf{K}$ and the rest parts are zero. $\mathbf{I}_k$ is a $N_k \times N_k$ diagonal matrix. $(\mathbf{I}_k - \mathbf{A}_k^{\phi})^T(\mathbf{I}_k - \mathbf{A}_k^{\phi})$ preserves locality of nearby points with same class label in the embedding space if they are close in original space during the unfolding process of nonlinear structures.

In this context, minimizing a objective $\mathbf{w}^T\mathbf{S}_w^{\phi}\mathbf{w}$ means to find a $\mathbf{w}$ that keeps the local geometry of within-class data as much as possible, so we integrate it to the *C*-SVM and define the primal problem as

$$\min_{\mathbf{w},b}\left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}L_c(\mathbf{x}_i) + \frac{\eta}{2N}\mathbf{w}^T\mathbf{S}_w^{\phi}\mathbf{w}\right), \tag{10}$$

where hinge loss $L_c(\mathbf{x}_i) = \max\left(0, 1 - y_i(\mathbf{w}^T\phi(\mathbf{x}_i)_{kpca} + b)\right)$, $C \geq 0$, and $\eta \geq 0$ stands for a trade-off parameter to balance the penalties of within-class neighborhood preserving and maximum margin of the decision hyperplane in *C*-SVM. If $\eta = 0$, the model will degrade to *C*-SVM where the within-class neighborhood preserved regularization does not work anymore.

Problem (10) is equivalent to the following formulation

$$\min_{\bar{\mathbf{w}},b}\left(\frac{1}{2}\bar{\mathbf{w}}^T\bar{\mathbf{w}} + C\sum_{i=1}^{N}L_c(\bar{\mathbf{x}}_i)\right), \tag{11}$$

where $\bar{\mathbf{w}} = \mathbf{S}^{1/2}\mathbf{w}$, $\bar{\mathbf{x}}_i = \mathbf{S}^{-1/2}\mathbf{W}_{kpca}^T\mathbf{K}(\mathbf{x}_i, \cdot)$ and $\mathbf{S} = \mathbf{I} + \eta/2N\mathbf{S}_w^{\phi}$. Thus it can be solved in standard *C*-SVM framework. We can simplify these computations via SVD (Singular Value Decomposition) technique to obtain $\mathbf{S}^{\pm 1/2}$ because $\mathbf{S}$ is a real symmetric matrix [48]. Assume the optimal result of problem (10) is $(\alpha^*, \mathbf{w}^*, b^*)$, the decision hyperplane becomes

$$g(\mathbf{x}) = sgn\left(\sum_{i=1}^{N}\alpha_i^* y_i\left(\mathbf{K}(\mathbf{x}_i, :)^T\mathbf{W}_{kpca}\mathbf{S}^{-1}\mathbf{W}_{kpca}^T\mathbf{K}(\mathbf{x}, :)\right) + b^*\right). \tag{12}$$

2.5. MIMO Within-Class Neighborhood Preserved ε-SVR Bounding Box Regression

Once the samples are classified as positive class using within-class neighborhood preserved *C*-SVC, they can be regarded as myocardium. However, their positions will not be completely overlapped with the true myocardium, due to the error of supervoxel merging and the classifier performance. To alleviate this side effect, we adopt bounding box regression technique to refine the proposal box.

Suppose (x, y, w, h) indicate the horizontal and vertical center coordinates of a detection box and its width and height, respectively, x_a and x_p are the center coordinates x of a detected box (called anchor) and a refined box (called proposal), respectively (the same applies to y, w and h), then a bounding box regression vector $\mathbf{z} \in \mathbb{R}^4$ could be presented by parameterizing the transformation between the anchor and the proposal (that is, the bounding box that seems to enclose a true myocardium), i.e., $z_1 = (x_p - x_a)/w_a$, $z_2 = (y_p - y_a)/h_a$, $z_3 = w_p/w_a$, $z_4 = h_p/h_a$. When $\mathbf{z}$ is learned from the samples and the ground truth, the anchor can be transformed into a refined proposal box. In this paper, we regard it as a multiple input multiple output (MIMO) regression problem and propose a MIMO within-class neighborhood preserved ε-support vector regression (SVR) to solve it. Different from the linear regression in RCNN [8] and Faster RCNN [9,10], the multidimensional regression will help

to exploit the dependencies in the channel and will make each estimate less vulnerable to the added noise. Treating all the channel paths together will allow to accurately estimate each of them when only scarce data is available.

The traditional solution of MIMO problem is splitting multi-dimensional output into multiple single-dimensional outputs, which means constructing an independent regression model for each output dimension. Although this kind of method has simple implementation, it is computationally expensive and incapable of containing useful information among outputs. Another solution is multivariate statistical regression. However, this kind of method is sensitive to the changes of data so that it cannot be applied broadly. In present machine learning techniques, artificial neural network is the most common method to establish MIMO model. However, when facing small-scale sample problem, this method easily falls into local minimum and leads to over-fitting [49,50].

Suppose the elements in $\{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^N$ denote SSAE learned samples and their corresponding multiple output, our MIMO ε-SVR approach defines the problem

$$\min_{\mathbf{T},\mathbf{b}} \left(\frac{1}{2} \sum_{j=1}^{4} ||\mathbf{t}^j||^2 + C \sum_{i=1}^{N} L_r(u_i) \right), \tag{13}$$

where $\mathbf{T} = [\mathbf{t}^1, \cdots, \mathbf{t}^4]$, $\mathbf{b} = [b_1, \cdots, b_4]^T$, $\mathbf{e}_i = \mathbf{z}_i - \mathbf{T}^T \phi(\mathbf{x}_i)_{kpca} - \mathbf{b}$, $u_i = \sqrt{\mathbf{e}_i^T \mathbf{e}_i}$, the loss function is $L_r(u_i) = (u_i - \varepsilon)^2$ when $u_i \geq \varepsilon$ and 0 otherwise. We follow the nonlinear mapping in the previous subsection as $\phi(\mathbf{x}_i)_{kpca} = \mathbf{S}^{-1/2} \mathbf{W}_{kpca}^T \mathbf{K}(\mathbf{x}_i, \cdot)$ and $\mathbf{S} = \mathbf{I} + \eta/2N \mathbf{S}_w^{\phi}$.

By adopting the cost function $L_r(\cdot)$, MIMO ϵ-SVR is capable of finding the dependencies between outputs, and can take advantage of the information of all outputs to get a robust solution. As problem (13) cannot be solved straightforwardly, an iterative method was proposed to obtain a desired solution. By introducing a first-order Taylor expansion of cost function $L_r(\cdot)$, the objective of problem (13) will be approximated by the following objective

$$\min_{\mathbf{T},\mathbf{b}} \left(\frac{1}{2} \sum_{j=1}^{4} ||\mathbf{t}^j||^2 + C \sum_{i=1}^{N} a_i u_i^2 + const \right), \tag{14}$$

where $a_i = 2\gamma(1 - \varepsilon/u_i^k)$ when $u_i^k \geq \varepsilon$ or 0 otherwise, *const* is constant term which does not depend on $\mathbf{T}$ and $\mathbf{b}$, and the superscript k denotes k-th iteration. According to the Representer Theorem, the best solution of minimization of problem (14) in feature space can be expressed as $\mathbf{t}^j = \sum_i \phi(\mathbf{x}_i)_{kpca} \beta^j$, then the linear search algorithm can be readily expressed in terms of β^j. In fact, our model can be solved using the standard approach [49,50], just need to replace the mapping by our $\phi(\mathbf{x}_i)_{kpca}$.

Once β^j has been computed, for a new SSAE-learned vector $\mathbf{x}_i$ with (x_a, y_a, w_a, h_a), we can estimate the j-th output as

$$\hat{z}^j(\mathbf{x}_i) = \left(\sum_{i=1}^{N} \left(\mathbf{K}(\mathbf{x}_i, :)^T \mathbf{W}_{kpca} \mathbf{S}^{-1} \mathbf{W}_{kpca}^T \mathbf{K}(\mathbf{x}, :) \right) \beta^j \right), \tag{15}$$

so the position of the corresponding refined proposal is $(x_a + w_a \hat{z}^1, y_a + h_a \hat{z}^2, w_a \hat{z}^3, h_a \hat{z}^4)$. It is noted that this nonlinear regression technique is not conducted on all of the supervoxel regions, but just regions near to the ground truth by using intersection-over-union to measure their overlapping (e.g., larger than 0.6). Thus, this step can be regarded as the closing step of refinement.

2.6. Refinement

There is a sample imbalance problem in our detection model. Specifically, for each image in the training set, the intersection-over-union (IoU) index can be computed between the labeled location of the target and generated location by the model: $IoU = \frac{GTB \cap DB}{GTB \cup DB}$, where GTB and DB denote the ground truth of the target and detection result respectively. When IoU approaches to 1, the detection result overlaps the ground truth in more parts and vice versa. Generally, when constructing a training

set, a region is regarded as positive sample if its IoU exceeds a threshold (e.g., 0.5), and negative if IoU is less than another threshold (e.g., 0.3). But the number of negative samples in the training set is practically much larger than the positive samples because the parts occupied by the target on an image are often much smaller than those of the non-target. This imbalance has a remarkable impact on the training of the model and results in a false positive problem where many negative samples are misclassified as positive samples.

We employ the hard negative mining [24] strategy to solve this problem. After the deep model and the classifier are trained, the training set is sent to the model, and the classifier gives the probability that each region belongs to the positive sample. Then we pick out the samples with high scores and misclassified into positive samples and called them as "hard to share negative samples". From them we chooses those samples that are not only misclassified but also less than 0.1 in IoU as difficult negative samples, then feed them into the pre-trained model and conduct iterative fine-tuning. By this way, the training samples are balanced to improve the model's performance.

Furthermore, for each candidate region, the proposed C-SVM classifier will output its probability of each class. There are usually many candidates around the true myocardium region, so the final detected area should be inferred from these regions. In our experiments, we employ the non-maximum suppression (NMS) algorithm [51] to eliminate redundant (cross-repeated) candidates and find the best myocardium position. Since the overlap between the candidates is sometimes relatively large, it is necessary to remove these regions with higher IoU scores (e.g., larger than 0.3) between the overlapped area. Usually, only a small number of the most likely areas are remained after this processing.

3. Datasets and Evaluation Metrics

3.1. Cardiac MRI Dataset and Preprocessing

The heart MRI data used in our experiments are from the publicly available Cardiac Atlas Project (CAP) data set, a collaborative database created by multiple organizations [52]. The data set contains 83 patients with short axis cardiac MRI images, where the MRI scanners used to collect these images include Siemens (Avanto 1.5T, Espree 1.5T and Symphony 1.5T), Philips (Achieva 1.5T, 3.0T and Intera 1.5T), and GE (Signa 1.5T). Because of the different characters of acquisition equipments, the parameters of the heart MRI images in different patients are varying. The formation of the images were in presence of remarkable offset or distortion. The parameters of the typical short-axis cardiac MRI images are: thickness 6 mm, gap 4 mm or thickness 8 mm, gap 2 mm. The size of these images is between 192×156 and 512×512, and the resolution of each voxel is between $0.7 \times 0.7 \times 6$ and $2.0 \times 2.0 \times 10$ (in mm). The number of slices per patient image sequence is $8 \sim 17$, and each slice corresponds to approximately $18 \sim 35$ images (one cardiac cycle).

In spite of these varying changes across subjects and acquisition equipments, we took myocardium as a preprocessing of the subsequent steps such as registration or segmentation, so we did not adjust these images into a common image space with same spatial resolution. We only rearranged their orientation as RAI automatically and employed the N4 regularization algorithm [53] to correct the deviation field of each heart MRI image.

The myocardium region in each image was manually labeled by two well-trained students. Since this labeling was a non-trivial and challenging problem due to the projective nature of the data, fuzzy organ boundaries, and large anatomical variability, their results were carefully cross-checked and further checked by a radiologist to made a final result as gold standard for evaluation. For each subject, this processing took about 2 h.

3.2. Evaluation Metrics

Considering the limited size of the data set, we randomly divided 83 cardiac image sequences into three folds of approximately equal size (28, 28, and 27 subjects) for training, validation, and test. The three-fold cross-validation was taken as the evaluation prototype, that is, our experiments

were conducted three rounds, in each round we trained our model in the training set, adjusted the hyper-parameters in the validation set, and applied the model in the remaining test set for measuring its performance.

We employed true positive rate (or Tpr, sensitivity, Se, recall), positive predictive value (or Ppv, precision), F1, and area under the receiver operating characteristic curve (AUC) to measure the performance of the proposed method. Tpr (resp. specificity) is a measure of effectiveness in identifying regions with positive (resp. negative) classifications. Specifically, the chosen metrics are defined as $Tpr = tp/(tp + fn)$ and $Ppv = tp/(tp + fp)$, where tp, tn, fp and fn indicate the true positive (correctly identified regions), true negative (correctly identified background regions), false positive (incorrectly identified regions), and false negative (incorrectly identified background regions), respectively, and all the pixels are equally treated towards their bounding box without considering the tissue they depict. $F1$ is a harmonic mean of the *precision* and *recall* measures and can be used to measure the degree of similarity of the two sets. The expression is expressed by the following equation $F1 = 2Tpr \times Ppv/(Tpr + Ppv)$. The value of $F1$ is between [0–1] and the larger the value of $F1$, the more similar the two sets. $DR = 2(|A \cap B|)/(|A \cup B|)$, where A is the ground truth region, B indicates the detected region, and $|A \cap B|$ and $|A \cup B|$ denote the number of pixels in the intersected region and in the union region, respectively. All of accuracy and AUC and DR measure the overall detection performance.

Statistical analysis is performed as appropriate in order to evaluate the relative performance of different detection methods. Due to the relatively small number of images, $p < 0.05$ is considered to be statistically significant. All the experiments were carried out in MATLAB2016a with deep learning toolbox and parallel computing toolbox on a PC with an Intel Core i7-3770K CPU, 3.70GHz, GTX1070 GPU, and 16GB RAM. Since the negative samples are remarkably more than the positive ones, total feeding all samples into the SSAE network will make the model biased. To alleviate this, a batch training strategy was adopted.

4. Experimental Results and Discussion

The proposed detection method was evaluated from two aspects: effectiveness of the parts in the model, and the comparison with close related state-of-the-art detection methods. Also an experimental investigation was carried out in the next section on the parameter setting, i.e., the number of supervoxels, the size of proposed region, the structure of the SSAE network and the C and β in our SVM classifier and regressor.

4.1. Parameters Setting

4.1.1. Number of Supervoxels and Training Image Set Building

There are two parameters (the number of initial supervoxels M and the final number of merged supervoxels K) in the candidate region generation module. Since we intend to make the true blood pool of left ventricle locate in the merged supervoxels, we employed the Dice ratio (DR) index to measure the similarity between the true object and the supervoxel in the region of this object under different M and K. The search range of the initial supervoxel is $\{300, 500, 700, 900\}$ and that of the merged supervoxel is $\{50, 100, 150, 200\}$. Figure 4 presents the DR values corresponding to different parameters and it is shown that when the initial number is greater than 500 and merged number is greater than 100, the DR values are acceptable. The maximum DR reaches for $M = 500$ and $K = 100$, so we choose them as the parameters in the following experiments.

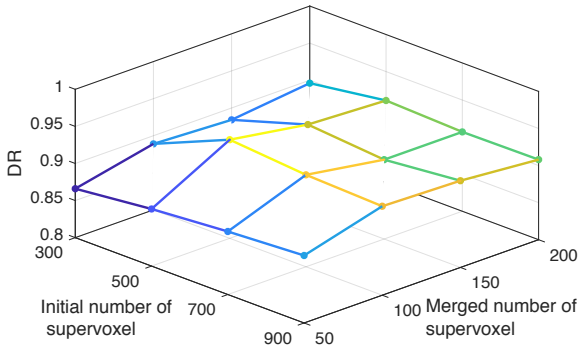

Figure 4. Mean Dice ratio on the training set using four different M and N values.

Once the merged number was settled, the positive samples were constructed as the bounding regions of the merged supervoxel that located in the real region of the left ventricle target on each training image; while the negative samples were those rest regions.

4.1.2. Size of Proposed Region and SSAE Training

The region proposals were scaled to a fixed size (i.e., $\tau \times \tau$) and sent to SSAE network for feature learning. The value of τ was determined through cross-validation to make the number of neurons in the SSAE input layer, specifically, the number of neuron in each encoding layer was hoped much less than that of the input neurons for achieving the sparse encoding, so we tested multiple values of τ in four kinds of three-hidden layer SSAEs with different numbers in each layer, i.e., $\{576, 400, 300, 200\}$ for $\tau = 24$; $\{1296, 750, 450, 250\}$ for $\tau = 36$; $\{2304, 1150, 600, 300\}$ for $\tau = 48$; and $\{3600, 1600, 750, 350\}$ for $\tau = 60$.

The models of four different SSAE structures were trained to adjust the optimal performance of each model in different SSAE super-parameters, including weight penalty factor λ, sparse penalty coefficient β, and coefficient factor ρ. The values of the three super-parameters were settled as $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, $\beta \in \{1, 0.6, 0.3, 0.1\}$, and $\rho \in \{0.01, 0.05, 0.1, 0.2\}$. Figure 5 presents the detection accuracy (F1) versus training time of the test model under four different τ values on the verification set. It can be seen from the figure that the overall trend is the higher the value of τ, the higher the detection accuracy of the model, along with the longer the training time. When τ reaches 60, the model obtains best test accuracy, so τ is set to 60. The values of three super-parameters are set to $\lambda = 10^{-2}$, $\beta = 0.3$, and $\rho = 0.2$.

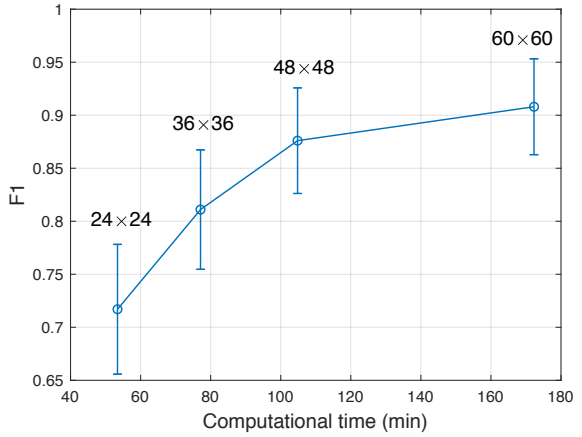

Figure 5. Detection accuracy ($F1$) on the verification set with model training time on the verification set using four different τ values.

4.1.3. Parameters of Within-class Neighborhood Preserved C-SVC and ε-SVR

Considering choosing the kernels and the parameters for the SVM-based methods is still an open problem, we adopted grid searching strategy to settle these parameters. The typical kernel used in our experiments is the Gaussian kernel, i.e., $\exp\left(-(\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v})/2\sigma^2\right)$ or $\exp\left(-r(\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v})\right)$, where σ controls the width of kernel while r is suitable for numerical searching. r should be larger than 0 in the sense of similarity. We selected r from $\{2^{-3}, 2^{-1}, 2^1, 2^3, 2^5\}$. For all C-SVM based methods, the common parameter is the slack variable C, we selected it from $\{2^{-3}, 2^{-1}, 2^1, 2^3, 2^5\}$; for the additional trade-off parameter η, we determined it from $\{2^{-3}, 2^{-1}, 2^1, 2^3\}$. To speed up the searching, the ranges were also restricted with respect to the data prior information. Figure 6 presents the average detection accuracies corresponding to different parameters. Overall speaking, smaller parameters are helpful for obtaining better accuracies, so we settle $C = 0.5$, $r = 2$, and $\eta = 2$ in all experiments.

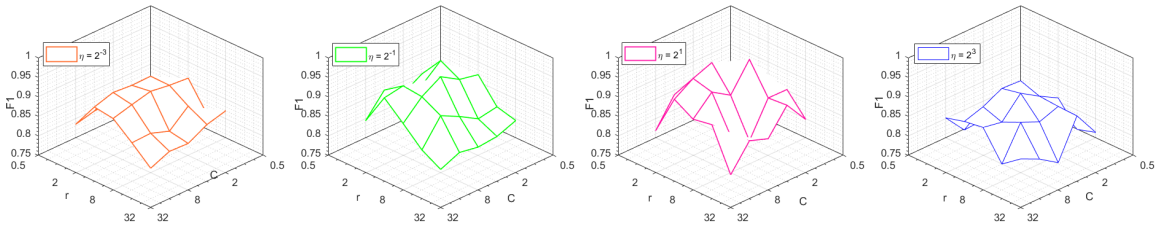

Figure 6. Detection accuracy ($F1$) on the verification set using different r, C, and η values.

Similarly, for parameters of within-class neighborhood preserved ε-SVR, we still adopted grid searching strategy to settle the parameters r in Gaussian kernel, the slack variable C, and the additional trade-off parameter η from the above ranges, and the parameter ε that sets the width of insensitivity zone of the regressors cost function was chosen from $\{0.01, 0.5, 1.0, 1.5, 2.0\}$. According to the best $F1$ accuracy, we settle $C = 0.5$, $r = 2$, $\eta = 2$, and $\varepsilon = 1.5$ in all experiments.

4.2. Validation of the Parts in Our Model

There are four main components in our model: structural similarity-enhanced supervoxel over-segmentation; deep SSAE feature learning; within-class neighborhood preserving-induced C-SVC and MIMO ε-SVR. To verify their roles in our model, we replaced each component into a state-of-the-art model and built five compared methods: (1) our model with SLIC that replaces our supervoxel over-segmentation; (2) our model with intensity feature that replaces SSAE learned feature; (3) our model with Softmax classifier that replaces the proposed classifier; (4) our model with C-SVM that replaces the proposed classifier; (5) our model with linear regression [8,9] that replaces the proposed regressor.

In order to objectively measure the performance of these five variations, the false positive rate and true positive rate of the detection results derived by different versions were calculated, by sweeping a threshold from 0 to 1 over the final classification output. The averaged results over them are plotted as receiver operating characteristics (ROC) curves in Figure 7. It is seen that our method with these components achieves the best performance. Also, when the within-class neighborhood preserved C-SVC is replaced by the standard C-SVM, the performance are better than those by the Softmax classifier. Furthermore, the over-segmentation and feature extraction methods are also essential for improving the overall accuracy. If the supervoxels were generated with SLIC, or only the intensity feature was adopted, the ROC curves increase much slower than others.

Table 1 shows the performances of different versions in detecting the myocardium. It shows that the proposed method achieves competitive results: the mean $F1$, Tpr, Ppv, and AUC are 0.924, 0.936, 0.916, and 0.891, respectively, remarkably higher than the proposed method with other modules. Statistical analysis shows that the performance of the proposed method is significantly higher with the SLIC, intensity, Softmax, and linear regression versions (in the level $p < 0.05$).

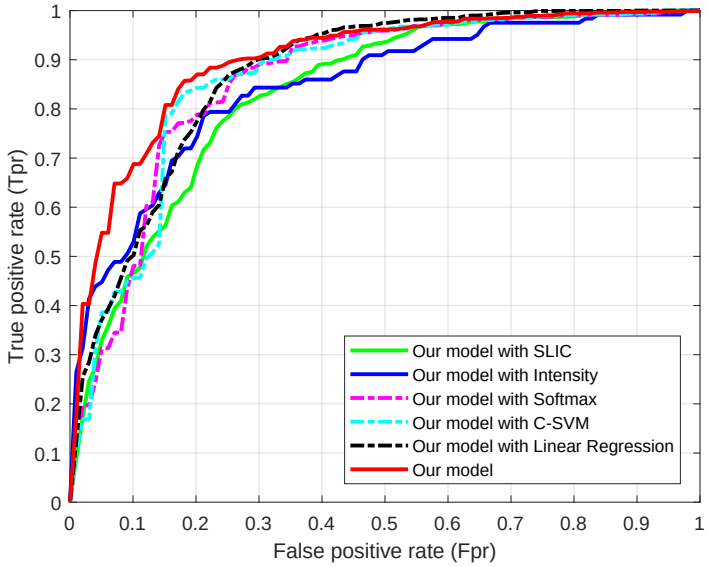

Figure 7. Receiver operating characteristics (ROC) curves of the proposed method with different modules.

Table 1. Performances (average ± standard deviation) of variations of our method.

Metric	Proposed Method with All Terms	Proposed Method with SLIC	Proposed Method with Intensity Feature	Proposed Method with Softmax	Proposed Method with *C*-SVM	Proposed Method with Linear Regression
*F*1	0.924 ± 0.034	0.852 ± 0.036	0.861 ± 0.038	0.878 ± 0.042	0.904 ± 0.035	0.898 ± 0.044
Tpr	0.936 ± 0.037	0.867 ± 0.048	0.884 ± 0.042	0.894 ± 0.040	0.915 ± 0.036	0.890 ± 0.038
Ppv	0.916 ± 0.028	0.838 ± 0.032	0.847 ± 0.037	0.866 ± 0.042	0.894 ± 0.039	0.885 ± 0.046
Area under ROC (*AUC*)	0.891 ± 0.031	0.824 ± 0.026	0.838 ± 0.032	0.851 ± 0.024	0.857 ± 0.030	0.862 ± 0.033

Figure 8 shows the detection results of different versions on three randomly selected cardiac MRI images in the test set, where red rectangles denote ground-truth and yellow ones denote results by compared methods. The larger overlapping means the corresponding method is better. As can be seen from the figure, our overall detection model is more robust, and the detection model based on *C*-SVM also achieves competitive results, which also demonstrates the strong ability of SSAE to learn the deep feature for classification, it can effectively distinguish different category of samples, thereby reducing the requirements of the follow-up classifiers.

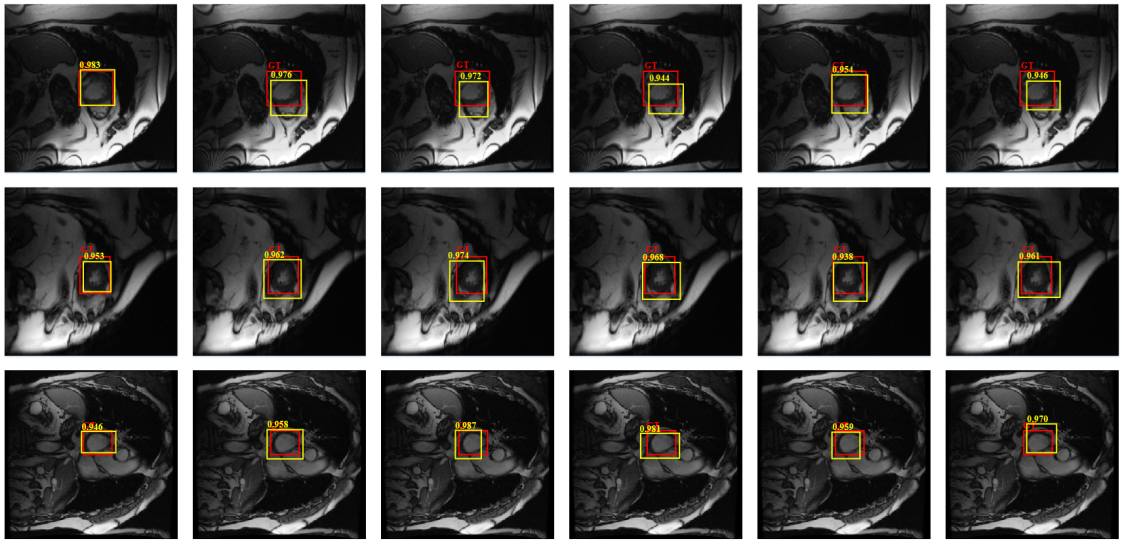

Figure 8. Detection results of our model with different parts, where red rectangles denote ground-truth and yellow ones denote results by compared methods. Each row from left to right: Our model; our model with SLIC; our model with intensity; our model with Softmax; our model with *C*-SVM; our model with linear regression, respectively.

4.3. Comparison with Related Methods

In this subsection, we carried out a comparative study between the proposed method and the state-of-the-art ones for the detection of myocardium over the cardiac datasets. Since the detection is based on the framework of region proposal and classification, such comparative study will help further explain its characteristic reported in the last section. To this end, five representative detection methods were selected: enhanced cascade detector (BCD) [6], RCNN [8], Faster RCNN [9], YOLOv3 [10,21], and a single-shot refinement neural network (RefineDet) [23].

BCD [6] is a well-known detection algorithm that was originally used in face detection and it trains the cascade AdaBoost to classify the regions that are generated by sliding windows and represented by Haar-like features. We took it as a representative of traditional detection methods.

RCNN [8] is a typical region proposal based convolutional neural network for object detection. This method firstly applies selective search method to generate around 2000 category-independent region proposals, then the features of each region proposal are extracted by a pre-trained convolutional model, finally the top-level features are classified by linear SVM. Faster RCNN [9] brings major improvements to traditional CNN by designing a region proposal network that extracts candidate areas instead of wasting time on selective search, which significantly accelerates the detection. On the other side, YOLOv3 [21] is an improved version of the state-of-the-art, real-time YOLOv2 [10] that applies a single neural network to the full image. The network divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities. We take it as a representative of region proposal independent neural network detection method. RefineDet [23] is a recent single-shot based detector that consists of the anchor refinement module and object detection module to achieve better accuracy than two-stage methods and maintains comparable efficiency of one-stage methods. We took the RefineDet320+ and VGG-16 net as the training model.

The average results over the compared methods are plotted as receiver operating characteristics (ROC) curves in Figure 9. It can be seen that our method consistently outperforms its competitors RCNN and Faster RCNN and it is also competitive to recent advancements YOLOv3 and RefineDet. Overall it achieves the best performance, while BCD is the worst among these methods. Furthermore, Faster RCNN performs similar to our method and outperforms RCNN.

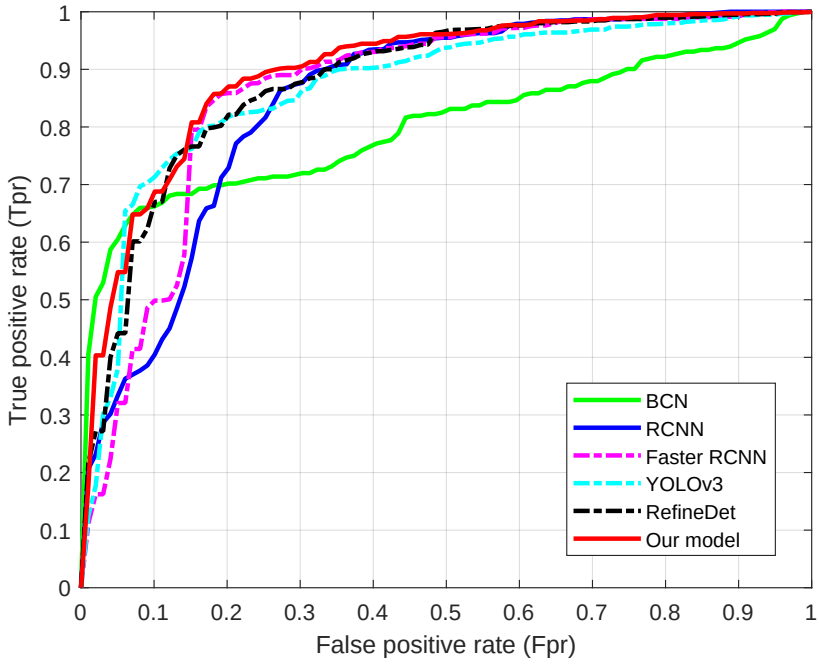

Figure 9. Receiver operating characteristics (ROC) curves of the compared methods.

The average accuracies of the detection based on the six methods are shown in Table 2. It can be seen from the table that the accuracy of the proposed algorithm is superior to that of the other five algorithms, the worst performance is BCD and Faster RCNN obtains the best accuracy among the convolutional neural network based methods.

Table 2. Performances (average ± standard deviation) of six compared methods.

Metric	Proposed Method	BCD	RCNN	Faster RCNN	YOLOv3	RefineDet
$F1$	0.924 ± 0.034	0.801 ± 0.092	0.870 ± 0.061	0.896 ± 0.058	0.878 ± 0.065	0.914 ± 0.046
Tpr	0.936 ± 0.037	0.805 ± 0.097	0.877 ± 0.062	0.908 ± 0.056	0.892 ± 0.062	0.918 ± 0.041
Ppv	0.916 ± 0.028	0.798 ± 0.103	0.863 ± 0.069	0.874 ± 0.062	0.862 ± 0.060	0.898 ± 0.045
Area under ROC (AUC)	0.891 ± 0.031	0.798 ± 0.026	0.858 ± 0.037	0.872 ± 0.025	0.870 ± 0.032	0.875 ± 0.036

Figure 10 shows the detection results of three randomly selected images in the testing set. In each row, red rectangles denote ground-truth and yellow ones denote results by compared methods. The values on the top of the bounding box is the maximal probability outputs by each method. The results show that our method can detect various myocardium with high quality, in most cases, the overlap between the detection of the final results and the target real area is higher, which is benefited from the combined candidate region generation, classification and regression algorithm. Faster RCNN and RefineDet also performs well in these images, except some small difference in overlapping. YOLOv3 also locates the accurate object in most images, but was disturbed by complicated surrounding tissues. In comparison, BCD is the worst detector and most resulting locations are low quality.

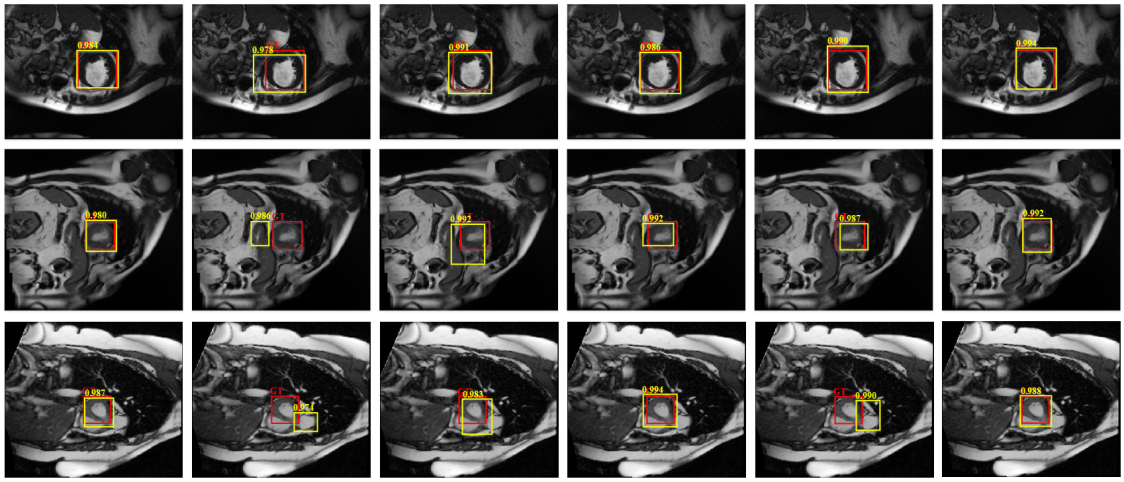

Figure 10. Examples of detection results, where red rectangles denote ground-truth and yellow ones denote results by compared methods. Each row from left to right: Our model; BCD; RCNN; Faster RCNN; YOLOv3, and RefineDet, respectively.

4.4. Discussion

4.4.1. Detection Performance

With regard to myocardium detection performance, the $F1$ and AUC measures indicate our method achieves a higher performance than previous studies using hand-crafted features such as the HOG or intensity in BCD method. They also tell that it is necessary to design a specific detection model towards a specific medical object. As we know, RCNN, Faster RCNN, YOLOv3, and RefineDet are state-of-the-arts in object detection, however, they are designed for general multiple objects detection and their advantages aren't thoroughly embodied for myocardium detection in our task. The $F1, Tpr, Ppv, AUC$ measures of our method are higher 1.0%, 1.8%, 1.8%, 1.6% than the best of these methods (RefineDet). Statistical analysis show that these four metrics of our method significantly outperforms Faster RCNN (in the level $p < 0.05$). The AUC and Ppv of our model significantly outperforms Fast YOLOv3 and RefineDet (in the level $p < 0.05$).

Our results indicate the strong classifiers such as *C*-SVM are helpful for achieving higher performance than Softmax. In our proposed *C*-SVM classifier detection model, SSAE and *C*-SVM training were carried out independently, that is, we first conducted SSAE unsupervised training, and then used the learned features to train the proposed *C*-SVM classifier. According to Softmax classifier, SSAE and Softmax training can be combined to form a whole part, that is, we first conducted SSAE unsupervised pre-training, and then connected SSAE and Softmax. The classification error of Softmax can be propagated back to SSAE, and the detection effect based on Softmax classifier is logically consistent. However, as pointed by RCNN [8], Softmax does not outperform SVM for classifying objects. Our experimental results also support this conclusion. Nevertheless, it is necessary to use the loss function of SVM to design error back propagation approach to the SSAE network in our next work.

With regard to region detection, although the myocardium is an ellipsoid-like tissue, the two-dimensional image of the myocardium varies in size and shape due to variations in motion and acquisition. For this reason, using only a single fixed-sized window is insufficient to detect myocardium regions with various shapes. Although a fixed-sized window has been used in RCNN, the detected myocardium is resized by a fixed-sized bounding box, regardless of its shape. On the other hand, our approach similarly feeds a fixed-sized window to SSAE network; however, inside SSAE, varying bounding boxes are possibly more suitable for the shape of myocardium by scanning the input image with multiple supervoxels with different scales and aspect ratios. To make use of SSAE, we set a slightly larger window to include multiple myocardium and background in the four corners. This was not optimal because SSAE learns the region of myocardium and background at the same time, and it is better to exclude various backgrounds in an input image to learn positive regions. This will be further investigated in our future work.

Furthermore, the structure prior information hidden in medical images is useful for accurate myocardium detection. For example, we proposed the structure similarity induced supervoxel instead of simple intensity similarity based supervoxel; we also proposed to incorporate the within-class neighborhood preserved scatter matrix to standard *C*-SVM classifier, which remarkably improves the overall performance of our model. As we can find from the first experiment that evaluated the major parts of our model, the consideration of structure prior information enhanced the distinctiveness of myocardium object from the complex background.

The detection model provides a good detection effect for most of the heart MRI images in the CAP data set, but there are still some failures in our experiments. Figure 11 presents some error detection. For the sake of comparison, two images of each column in the figure come from the same patient's heart MRI image sequence, which shows that the place where false detection occurred. The errors mainly appear at the head and tail of the heart image sequence. At these places, the sizes of the left ventricle tissue are very small, the shapes are not obvious, and the right ventricle occurred with more adhesion to left ventricle. When the left ventricle with the normal form (as shown in the first line), the accuracy of outer frame is almost the same as that of the real bounding box.

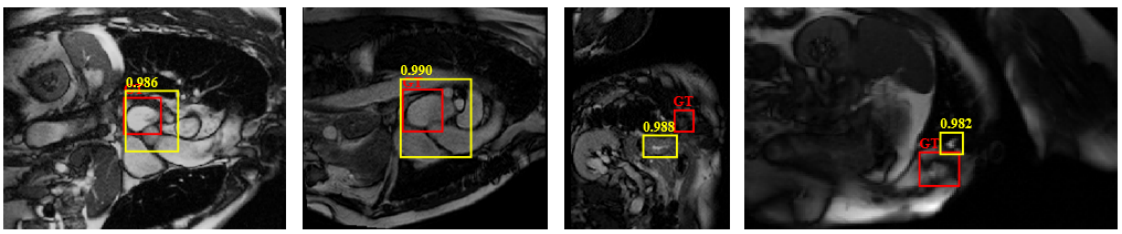

Figure 11. Examples of failure detection results by our method, where red rectangles denote ground-truth and yellow ones denote our results.

4.4.2. Processing Speed

The four parts, i.e., the supervoxel-based region proposal, the SSAE feature learning, and the within-class neighborhood preserved *C*-SVM, and ε-SVR, in our model took different time in training.

Compared with thirteen minutes to train the SVM model and one minute to generate region proposals, the SSAE took much longer time and it took about three hours. Because our approach feeds relatively large-size images to SSAE, it requires the number of mini-batches to be reduced to one because of the size of the GPU memory, leading to a slow learning rate for stable learning. Therefore, it is necessary to increase the number of iterations to train the network sufficiently. Fortunately, our SSAE is not a deeper structure that contributes to the long training time, like the VGG-16 in some CNN based models. Once the training is finished, our model averagely took less than one minute to extract the myocardium from a cardiac image with the trained model. This performance seems to be comparable in medical applications. This high-throughput approach may have some advantages in practical usage in hospitals and laboratories to assist pathologists in their daily tasks.

4.4.3. Limitations of the Work

There are several limitations that need to be addressed. Firstly, this work, like most neural network-based cardiac MR image analysis studies [4,29,30], suffers from the restriction of available ground truth data to a limited number of cardiovascular disease diagnoses, such as pulmonary hypertension, congenital heart disease, coronary heart disease, and dysplasia. Therefore, results of this study can only show performance on limited set of patients. The number ot layers in SSAE learning was also restricted. Besides, currently available data largely consists of short-axis image where boundaries between the blood pool and myocardium are more or less clearly visible. More challenging image such as long-axis image with large spatial interval sampling are not part of the current sets and should be researched on in future work.

More importantly, this study focused on the relation of left ventricle and its myocardium, and didn't consider the relation of other cardiac structures (such as right ventricle and its myocardium, left atrium, right atrium). In fact, these structures have strong anatomical prior that can guide the accurate localization of myocardium and left ventricle. Future studies may gain insight from the recent advance on the CNN-based relation modeling among detected objects [18], the CNN-based iterative localization refinement [19], and explore the multiple objects detection by extending the proposed framework.

5. Conclusions

In cardiac MR image analysis, left ventricle and myocardium detection is often used as a prerequisite step, which plays a key role in the successive steps such as image registration and segmentation. This paper has presented a new efficient detection approach to myocardium structures in cardiac MR images through an enhanced region proposal-based model. The model first proposed a structural similarity-enhanced supervoxel over-segmentation and hierarchical clustering approach to extract candidate regions; then, the deep features were learned by SSAE network; furthermore, the learned features are classified by a within-class neighborhood preserved C-SVC, and during the refinement, the bounding boxes are adjusted by a multiple-intput multiple-output within-class neighborhood preserved ε-SVR regression and hard negative sample mining technique. Different parts in our model were also tested and prediction accuracies validated the advantage of proposed integration. Furthermore, comparative experiments demonstrated that the proposed model achieved a better detection accuracy on the publicly available dataset. The model does not require a large amount of training data and learns from coarsely annotated volumetric images (bounding-box masks). It can be potentially extended to similar object detection in other medical MR images.

Author Contributions: All authors have made substantial contributions to the study including conceptualization, methodology, algorithm design and experiments; writing—original draft preparation: Y.N.; writing—review and editing: Y.N., Q.L., X.W.

Funding: This research was funded by the Basic and Frontier Planning of CQ-CSTC (cstc2016jcyjA0317).

Acknowledgments: The authors would like to thank the Cardiac Atlas Project (CAP) that provides the whole complete data set with public domain license. The SLIC and over-segmentation programs came from OpenCV3.2 (https://opencv.org/). The experimental SSAE program came from the rasmusbergpalm-

DeepLearnToolbox-9faf641 toolkit (https://github.com/rasmusbergpalm/DeepLearnToolbox). The experimental SVM program came from libsvm-3.23 (http://www.csie.ntu.edu.tw/~cjlin/libsvm). The implementation of compared methods came from the codes link in corresponding reference papers. The authors also would like to thank the anonymous reviewers for their valuable comments, suggestions, and enlightenment.

Conflicts of Interest: The authors have no conflict of interest in the paper.

References

1. Xue, W.; Brahm, G.; Pandey, S.; Leung, S.; Li, S. Full left ventricle quantification via deep multitask relationships learning. *Med. Image Anal.* **2018**, *43*, 54–65. [CrossRef]
2. Khened, M.; Kollerathu, V.A.; Krishnamurthi, G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Med. Image Anal.* **2019**, *51*, 21–45. [CrossRef] [PubMed]
3. Mo, Y.; Liu, F.; McIlwraith, D.; Yang, G.; Zhang, J.; He, T.; Guo, Y. The Deep Poincaré Map: A Novel Approach for Left Ventricle Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer: Cham, Swetzerland, 2018. [CrossRef]
4. Vigneault, D.M.; Xie, W.; Ho, C.Y.; Bluemke, D.A.; Noble, J.A. Ω-Net (Omega-Net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks. *Med. Image Anal.* **2018**, *48*, 95–106. [CrossRef] [PubMed]
5. Larroza, A.; Lopez-Lereu, M.P.; Monmeneu, J.V.; Bodi, V.; Moratal, D. Texture analysis for infarcted myocardium detection on delayed enhancement MRI. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, Australia, 18–21 April 2017; pp. 1066–1069.
6. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
7. Criminisi, A.; Shotton, J.; Bucciarelli, S. Decision forests with long-range spatial context for organ localization in CT volumes. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2009, London, UK, 20–24 September 2009; pp. 69–80.
8. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
10. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
11. Fareed, M.M.S.; Chun, Q.; Ahmed, G.; Murtaza, A.; Asif, M.R.; Fareed, M.Z. Appearance-Based Salient Regions Detection Using Side-Specific Dictionaries. *Sensors* **2019**, *19*, 421. [CrossRef] [PubMed]
12. Gao, L.; He, Y.; Sun, X.; Jia, X.; Zhang, B. Incorporating Negative Sample Training for Ship Detection Based on Deep Learning. *Sensors* **2019**, *19*, 684. [CrossRef]
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
14. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 1440–1448.
15. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
16. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 354–370.
17. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.

18. Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. Relation Networks for Object Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3588–3597.
19. Cheng, K.; Chen, Y.; Fang, W. Improved Object Detection With Iterative Localization Refinement in Convolutional Neural Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2261–2275. [CrossRef]
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
21. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767v1.
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016.
23. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-Shot Refinement Neural Network for Object Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4203–4212. [CrossRef]
24. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
25. Yan, Z.; Zhan, Y.; Peng, Z.; Liao, S.; Shinagawa, Y.; Metaxas, D.N.; Zhou, X.S. Bodypart recognition using multi-stage deep learning. In *Information Processing in Medical Imaging*; Ourselin, S., Alexander, D., Westin, C.F., Cardoso, M., Eds; Springer: Cham, Switzerland, 2015.
26. De Vos, B.D.; Wolterink, J.M.; de Jong, P.A.; Viergever, M.A.; Išgum, I. 2D image classification for 3D anatomy localization: Employing deep convolutional neural networks. In Proceedings of the SPIE Medical Imaging, San Diego, CA, USA, 27 February–3 March 2016.
27. Roth, H.R.; Lee, C.T.; Shin, H.C.; Seff, A.; Kim, L.; Yao, J.; Lu, L.; Summers, R.M. Anatomy-specific classification of medical images using deep convolutional nets. In Proceedings of the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, USA, 16–19 April 2015; pp. 101–104.
28. Luo, G.; An, R.; Wang, K.; Dong, S.; Zhang, H. A deep learning network for right ventricle segmentation in short-axis MRI. In Proceedings of the 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 11–14 September 2016; pp. 485–488.
29. Poudel, R.P.K.; Lamata, P.; Montana, G. Recurrent Fully Convolutional Neural Networks for Multi-slice MRI Cardiac Segmentation. In *Reconstruction, Segmentation, and Analysis of Medical Images*; Zuluaga, M.A., Bhatia, K., Kainz, B., Moghari, M.H., Pace, D.F., Eds.; Srpinger: Cham, Switzerland, 2016.
30. Tan, L.K.; Liew, Y.M.; Lim, E.; McLaughlin, R.A. Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine MR sequences. *Med. Image Anal.* **2017**, *39*, 78–86. [CrossRef] [PubMed]
31. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 391–405.
32. Fulkerson, B.; Vedaldi, A.; Soatto, S. Class segmentation and object localization with superpixel neighborhoods. In Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 670–677.
33. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]
34. He, S.; Lau, R.W.H.; Liu, W.; Zhe, H.; Yang, Q. SuperCNN: A Superpixelwise Convolutional Neural Network for Salient Object Detection. *Int. J. Comput. Vision* **2015**, *115*, 330–344. [CrossRef]
35. Stutz, D.; Hermans, A.; Leibe, B. Superpixels: An evaluation of the state-of-the-art. *Comput. Vision Image Underst.* **2018**, *166*, 1–27. [CrossRef]
36. Kovesi, P. Phase Congruency Detects Corners and Edges. In Proceedings of the 7th International Conference on Digital Image Computing: Techniques and Applications, Sydney, Australia, 10–12 December 2003; pp. 309–318.
37. Hinton, G.E.; Osindero, S.; Teh, Y.W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]

38. Yang, G.; Zhuang, X.; Khan, H.; Haldar, S.; Nyktari, E.; Ye, X.; Slabaugh, G.G.; Wong, T.; Mohiaddin, R.; Keegan, J.; et al. A fully automatic deep learning method for atrial scarring segmentation from late gadolinium-enhanced MRI images. In Proceedings of the 14th IEEE International Symposium on Biomedical Imaging, Melbourne, Australia, 18–21 April 2017; pp. 844–848. [CrossRef]
39. Yang, G.; Zhuang, X.; Khan, H.; Haldar, S.; Nyktari, E.; Ye, X.; Slabaugh, G.G.; Wong, T.; Mohiaddin, R.; Keegan, J.; et al. Segmenting Atrial Fibrosis from Late Gadolinium-Enhanced Cardiac MRI by Deep-Learned Features with Stacked Sparse Auto-Encoders. In Proceedings of the 21st Annual Conference on Medical Image Understanding and Analysis, Edinburgh, UK, 11–13 July 2017; pp. 195–206. [CrossRef]
40. Yang, G.; Zhuang, X.; Khan, H.; Haldar, S.; Nyktari, E.; Li, L.; Ye, X.; Slabaugh, G.G.; Wong, T.; Mohiaddin, R.; et al. Multi-atlas propagation based left atrium segmentation coupled with super-voxel based pulmonary veins delineation in late gadolinium-enhanced cardiac MRI. In Proceedings of the SPIE 2017 Medical Imaging, Orlando, FL, USA, 11–16 February 2017. [CrossRef]
41. Xu, J.; Xiang, L.; Liu, Q.; Gilmore, H.; Wu, J.; Tang, J.; Madabhushi, A. Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images. *IEEE Trans. Med. Imaging* **2016**, *35*, 119–130. [CrossRef] [PubMed]
42. Zhao, G.; Wang, X.; Niu, Y.; Tan, L.; Zhang, S. Segmenting Brain Tissues from Chinese Visible Human Dataset by Deep-Learned Features with Stacked Autoencoder. *BioMed Res. Int.* **2016**, *2016*, 5284586. [CrossRef]
43. Tao, C.; Pan, H.; Li, Y.; Zou, Z. Unsupervised Spectral–Spatial Feature Learning With Stacked Sparse Autoencoder for Hyperspectral Imagery Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2438–2442. [CrossRef]
44. Yan, Y.; Tan, Z.; Su, N.; Zhao, C. Building Extraction Based on an Optimized Stacked Sparse Autoencoder of Structure and Training Samples Using LIDAR DSM and Optical Images. *Sensors* **2017**, *17*, 1957. [CrossRef]
45. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Parallel Distrib. Process.* **1986**, *323*, 399–421. [CrossRef]
46. Ju, Y.; Guo, J.; Liu, S. A Deep Learning Method Combined Sparse Autoencoder with SVM. In Proceedings of the 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Xi'an, China, 17–19 September 2015; pp. 257–260.
47. Wang, X.; Niu, Y. Improved support vectors for classification through preserving neighborhood geometric structure constraint. *Opt. Eng.* **2011**, *50*, 087202.
48. Wang, X.; Niu, Y. New one-versus-all v-SVM solving intra-inter class imbalance with extended manifold regularization and localized relative maximum margin. *Neurocomputing* **2013**, *115*, 106–121. [CrossRef]
49. Fernández, M.S.; de Prado-Cumplido, M.; Arenas-García, J.; Pérez-Cruz, F. SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE Trans. Signal Process.* **2004**, *52*, 2298–2307. [CrossRef]
50. Mao, W.; Xu, J.; Wang, C.; Dong, L. A fast and robust model selection algorithm for multi-input multi-output support vector machine. *Neurocomputing* **2014**, *130*, 10–19. [CrossRef]
51. Wan, L.; Eigen, D.; Fergus, R. End-to-end integration of a convolution network, deformable parts model and non-maximum suppression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 851–859.
52. Fonseca, C.G.; Backhaus, M.; Bluemke, D.A.; Britten, R.D.; Chung, J.D.; Cowan, B.R.; Dinov, I.D.; Finn, J.P.; Hunter, P.J.; Kadish, A.H. The Cardiac Atlas Project—An imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics* **2011**, *27*, 2288–2295. [CrossRef]
53. Tustison, N.J.; Avants, B.B.; Cook, P.A.; Zheng, Y.; Egan, A.; Yushkevich, P.A.; Gee, J.C. N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging* **2010**, *29*, 1310–1320. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).