View Classification and Object Detection in Cardiac Ultrasound to Localize Valves via Deep Learning

Author(s) names withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2020

Abstract

Echocardiography provides an important tool for clinicians to observe the function of the heart in real time, at low cost, and without harmful radiation. Automated localization and classification of heart valves enables automatic extraction of quantities associated with heart mechanical function and related blood flow measurements. We propose a machine learning pipeline that uses deep neural networks for separate classification and localization steps. As the first step in the pipeline, we apply view classification to echocardiograms with ten unique anatomic views of the heart. In the second step, we apply deep learning-based object detection to both localize and identify the valves. Image segmentation based object detection in echocardiography has been shown in many earlier studies but, to the best of our knowledge, this is the first study that predicts the bounding boxes around the valves along with classification from 2D ultrasound images with the help of deep neural networks. Our object detection experiments suggest that it is possible to localize and identify multiple valves precisely.

Keywords: ultrasound, echocardiagram, view classification, object detection, valve localization, deep learning.

1. Introduction

Echocardiography (diagnostic cardiac ultrasound imaging) is routinely used to visualize the chambers and valves of the heart. Typically, it is combined with Doppler ultrasound to evaluate blood flow through valves and within chambers. High frequency sound waves are transmitted into the body, and the received echoes from tissue are processed to produce both 2D images, and blood flow velocity estimates.

This study focuses on heart valves which are critical components of the heart, namely the valves that control circulation between the chambers and aorta: Tricuspid Valve (TV), Mitral Valve (MV) and Aortic Valve (AV). TV is located in the right side of the heart, AV and MV is in the left side. TV and MV are called atrioventricular valves, since they connect the atria to the ventricles. In an ideal heart, blood flows through the both valves during diastole with contraction of the corresponding atrium and both close during systole with contraction of the corresponding ventricle to prevent regurgitation of blood from the ventricle to the atria. On the other hand, the AV is responsible for controlling the blood circulation between the left ventricle and aorta, which is the main artery supplying oxygenated blood to the circulatory system. In the case of a heart with pathology, the blood may flow backwards through the valve if the valve does not close completely (regurgitation or insufficiency of the valve). MV and AV regurgitation affect more than 200.000 people

per year in United States. Another significant valve abnormality is stenosis where the valve flaps become stiff resulting in narrowed valve openings and reduced blood flow. Any significant valve insufficiency and abnormality can affect the quality of daily life, and may require significant treatment procedures. Untreated pathologies can result in enlargement of heart, heart rhythm problems (arrhythmia), heart failure or even death.

Classification and localization of anatomy are key enabling technologies that open up doors to many solutions for Ultrasound techs as well as clinicians. Training and placement guidance for new and/or inexperienced users is an application that would tremendously benefit from these technologies. Successful localization and identification of valves would allow automatic highlighting and enhancing these organs in the image to make accurate measurements, guide procedures, place devices, etc.

1.1. Related Work

Deep neural networks are beginning to assist Echocardiagram image analysis. Automatic view classification is a popular application area since it is a basis for many other applications. Machine learning and image processing techniques for ultrasound image classification have been explored by many papers. Here, we focus only on those papers using deep neural networks. In (Gao et al., 2017), two convolutional neural networks (CNNs) are combined to classify eight different views, namely Apical 2, 3, 4, 5 chamber, parasternal long axis (PLAX), parasternal short axis at aortic valve (PSAX-AoV), PSAX of papillary (PSAX-LV) and PSAX at mitral valve (PSAX-MV). In addition to brightness mode (B-mode) images, the temporal acceleration images are processed by a different network and the results are fused to obtain a final decision. This combination provided an average of 92% accuracy for all views, with the lowest accuracy (71.4%) in Apical 5 (mostly mixed with Apical 3 class) . Another study on view classification was done by (Madani et al., 2018). In addition to the classes listed above, they included PLAX RVIF, subcostal4C, subcostal inferior vena-cava, subcostal aorta, suprasternal aorta. Using the VGG-16 network(Simonyan and Zisserman, 2014), they achieved 97.8% overall accuracy in 15 different views with the accuracy for PLAX RVIF, specifically 86% in video and 72% on still images.This exceeds the prediction accuracy of a board-certified echo cardiographer. A more extended classification study was performed by (Zhang et al., 2018), which also included subclasses of certain views. For example, in addition to the typical Apical 2, the study included same view with occluded left atrium, and occluded left ventricle. Due to the high correlations between classes, their average accuracy was 84% on 23 views. However, if the results are considered in terms of broad classes such as PLAX, the accuracy they achieved was around 96%. The classification network used was the 15 layer VGG network (Simonyan and Zisserman, 2014). In addition to the view classification, they used deep networks also for image segmentation of cardiac chambers and also disease classification. However, image segmentation and disease detection are outside the scope of this paper.

Several algebraic, signal processing and machine learning techniques have been proposed for tracking cardiac valves. In (Dukler et al., 2018), a non-negative matrix approximation approach has been used to detect and track the mitral valve in Apical 4 chamber views. This algorithmic approach does not require any labeling of the data but is limited to detecting the mitral valve only. Reference (Voigt et al., 2015) uses a machine learning approach for

real-time tracking of mitral valve in 3D images for inteventional guidance. Their technique relies on the box estimator, based on *marginal space learning (MSL) approach* (Zheng et al., 2009) that predicts the presence of the MV location, orientation and scale in 3D images. MSL is three-step detector that applies probabilistic boosted-tree based classifiers multiple times to estimate different parameters. None of these studies uses deep neural networks.

Convolutional networks in conjunction with object detection are commonly used in medical imaging for localization and segmentation of anatomical structures and organs. For example, (de Vos et al., 2017) uses a specialized convolutional network called BoBNet (a variation of VGG network(Simonyan and Zisserman, 2014)) to predict bounding boxes around organs such as liver, heart, aorta applied to 3D computerized tomography images (CT) images. On the other hand, there is a lot of effort in fetal ultrasound imaging to identify the imaging plane and detect the structures being imaged. For example, in (Huang et al., 2017b) and (Sundaresan et al., 2017), CNN based techniques are proposed to automatically localize a fetal heart. Another CNN based technique (Baumgartner et al., 2017) is able to detect fetal standard planes and localize structures such as brain, spine, kidneys, lips, femur, etc. from 2D ultrasound images. CNNs are commonly used to segment the cardiac valve leaflets. For example in (Costa et al., 2019) uses CNNs to segment the mitral valve leaflets but this is limited to two views (Apical 4 and PLAX).

In this paper, we illustrate how deep neural networks developed for the object detection problem perform on cardiac ultrasound images to localize the valves in different cardiac views. In the first section, we explain the image preprocessing and view classification applied as the initial step. Then, in the second section, we concentrate on the annotated data specific to object detection and mention the selected network for training. In the final section, we show the results of our experiments. To the best of author's knowledge, this is the most compherensive study which provides and end-to-end machine learning pipeline to localize the cardiac valves using object detection networks in many different cardiac views.

2. Materials and Methods

2.1. View Classification

View Classification of selected echocardiogram views is the first stage in our proposed machine learning pipeline. Views classified are as follows: Apical 2, Apical 3, Apical 4, Apical 5, Parasternal-long-axis (PLAX), PLAX-RVinflow (PLAX-RVIF), PLAX-RVoutflow (PLAX-RVOT), Parasternal-short-axis (PSAX)[1], PSAX at the aortic valve level (PSAX-AoV), Subcostal of four-chamber, and Noise. The noise images for training are created by capturing images where the ultrasound probe is in contact with air, or in contact with ultrasound coupling gel only. An example image for each cardiac image class is shown in Figure 1c.

1. PSAX view can be from Left Ventricle level or Mitral Valve. We observed that those two views look very similar and object detection to isolate the valve would produce similar results, thus we merged these two sub-classes.

2.1.1. Annotateted Data

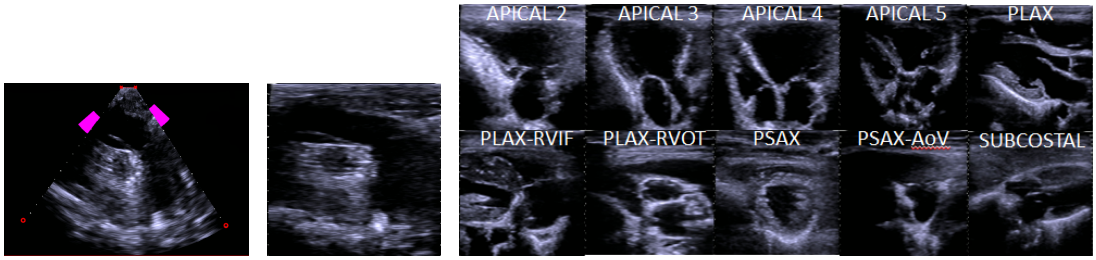
The complete dataset includes 11,150 B-mode clips of 1-5 heartbeats from Acuson SC2000, Siemens Cardiac Ultrasound system (Mountain View, CA, USA). All the clips are anonymized to remove any patient specific information. To ensure that the view classification testing is valid, care was taken to ensure that B-Mode clips from the same patient do not appear in both the training and the validation/test data to eliminate bias. Because the data is anonymous, we used acquisition date and time that was contained in the DICOM header files. We numbered the clips with a separate patient ID if there is 30 minutes gap between two closest acquisitions. This method may result in identifying two separate patients as the same person (if two studies were done back-to-back). However, it has low probability of splitting the images from a single patient into two parts. With this method, we identified the total number of subjects to be at least 525. We partitioned 60% of patients for training, 20% for validation and 20% for testing. The distribution of overall data is shown in Figure A.

2.1.2. Preprocessing

The images were pre-processed in preparation for training. First, we divided the clips into frames (frame rate was 50-70 frames per second) and randomly selected 10 frames per heartbeat up to a maximum of 30 frames per clip. This allowed us to have 126,731 images for training, 38,148 for validation and 35,932 for testing. In typical ultrasound images, anatomical structures are shown in a polar coordinate system within a trapezoid-shaped area as shown in Figure 1a. There also exists some text related to system, acquisition or patient related information on the left and right side of rectangular images. In order to provide only necessary information to our network, the data is converted from the display grid (the trapezoid shape) to a Cartesian grid in the first pre-processing step. The coordinates of the image region is provided in the DICOM header acquired by the SC2000 system. Then the image within the trapezoid is transformed into Cartesian coordinate system via linear interpolation to give us a converted image as shown in Figure 1b. The images were resized to 256×256. The mean calculated from the training set images is extracted from each image. The grayscale images (maximum value of 255) were normalized to have values between 0-1 before feeding into the network.

2.1.3. Classification Network

We adopted the **InceptionV3** network as described in (Szegedy et al., 2016). In this network, each convolutional layer follows batch normalization and activation units. Naive Inception modules include combinations of 1×1, 3×3, 5×5 convolutional layers plus a 3×3 pooling layer. The output of all those layers are then concatenated to create the input to the next layer. In InceptionV3, several factorization and dimensionality reduction techniques are used in Inception layers to increase computational efficiency. For example, most convolutional layers are preceded by a 1×1 block to reduce dimension along depth. InceptionV3 starts with three cascaded convolutional layers, whose dimensions are $3 \times 3 \times 32$ (stride 2), $3 \times 3 \times 32$ (stride 1), $3 \times 3 \times 64$ (stride 1), respectively. It is followed with a 3×3 (stride 2) maximum pooling layer. Then, two cascaded convolutional layers again with filter sizes $1 \times 1 \times 80$ (stride 1) and $3 \times 3 \times 192$ (stride 1); this again is followed with a maximum



(a) (b) (c)

Figure 1: An example input image (a) before and (b) after pre-processing; (c) distribution of DICOM clips in overall data

pooling layer with 3×3 (stride 2). After this, 15 different Inception modules that use all the computational techniques explained above follow. For an input size $299 \times 299 \times 3$, the output size at the end of the complete network becomes $8 \times 8 \times 2048$.

In our experiments, we used Python 3.6.5 from Anaconda (Austin, TX, USA), Keras 2.2.2, Keras Applications 1.0.5 and Keras Preprocessing 1.0.3 packages. Keras was set to work in Tensorflow (Abadi et al., 2016) backend. We used tf-nightly-gpu (version 1.10.0) for Tensorflow (Google Inc, Mountain View, CA, USA). Training was run on a 64 bit Windows 10 system with Intel®Xeon®CPU E5-2640 processor, a single Nvidia®Geforce Titan X (12GB) GPU card and 32GB RAM. The training took around 12 hours for 20 epochs with a training batch size 64 ($\approx 40,000$ iterations). After each epoch, prediction on a validation set was performed and the model with lowest validation loss among 20 iterations was used to determine the final model. The loss function was the categorical cross-entropy function, used along with the Adadelta optimizer (learning rate=1.0, rho=0.95, epsilon=1e-8, decay=0). The class with maximum score was used as the final prediction value. We also applied random data augmentation with zoom range up to 15%, shear range up to 3%, height and width shift ranges up to 15% , rotation angles up to 10 degrees, contrast range from -100 to 40. Please see the Results section for the classification results.

2.2. Object Detection

In order to localize the valves, we applied object detection training for all Apical, PLAX and PLAX-RVIF views separately[2]. We used Tensorflow's Object Detection API (Huang et al., 2017a), which includes implementation of well-known deep learning based networks (SSD, RFCN, Faster-RCNN) and the tools for easy training and testing. The system we used for training was the same system and same packages/software mentioned in the View Classification section.

The number of classes for identification within the object detection portion of our pipeline depends on the view classification result obtained in in the first stage of the pipeline. Apical 2 consists of the MV, Apical 3 has MV and AV, Apical 4 has MV and TV. In Apical 5 both MV and TV may appear partially but the region of clinical interest is left ventricle

2. The annotations for the other views are not yet complete.

outflow tract (LVOT) location just above AV. Additionally, PLAX consists of MV and AV; and PLAX-RVIF consists of TV. Thus, we train object detection networks separately for each cardiac view.

2.2.1. Annotated data

As a basis for object detection, the B-Mode DICOM clips were annotated. Only one frame within a B-Mode DICOM clip was annotated, however we used the derived bounding box as ground truth for neighboring frames in the clip as well, assuming that there is no other motion than heart-motion. Annotations were done on the frame in the heart cycle where the valve is completely closed. The annotation specifies three coordinate locations within the image: the center point where valve flaps touch when fully closed, and the left and right points where the valve connects to heart wall tissue. Since the annotation did not include top and bottom coordinates to define a ground truth bounding box, we selected a fixed height for all bounding boxes which is large enough to encompass the valve when it is fully open.

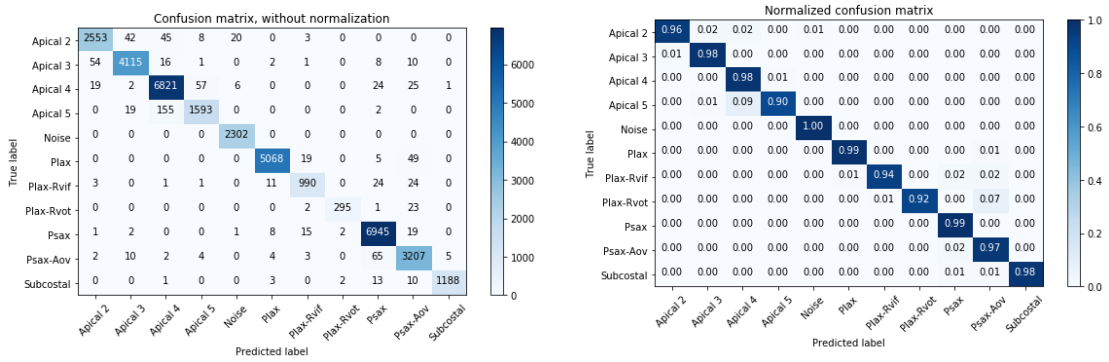
2.2.2. Object Detection Network

The meta-structure, we adopted was Faster R-CNN (Ren et al., 2015), consisting of two stages. The first stage is called *region proposal network (RPN)*, where the features are extracted from the intermediate layers of the networks such as Inception, ResNet or VGG. These features are given to a *region proposal generator* that outputs the bounding box coordinates and object-ness scores of fixed number of regions (e.g.300). In the second stage, a cropped set of sub-images created using the region proposals is fed to the remainder of the feature extractor network to output the predicted class and refined bounding box coordinates. Since this operation is done separately for each proposed region, the speed of the Faster R-CNN is highly dependent on the number of region proposals selected. Details about the loss functions and speed/accuracy comparisons of different feature extractors and meta-structures can be found in (Huang et al., 2017a).

We specifically used **Faster-RCNN with ResNet101** (He et al., 2016). This network has been shown to be slower than other deep learning based networks such as SSD and R-FCN but provides more accurate results (Huang et al., 2017a) based on experiments done on the Microsoft COCO dataset (Lin et al., 2014). The speed and accuracy rates of this network are highly dependent on parameters such as input image size and the number of bounding box proposals. In our applications we used low-resolution images (256×256) but selected the number of proposals as 300.

We started with a pre-trained faster-RCNN network on the COCO dataset (the checkpoint was downloaded from the Model Zoo website[3]). The data augmentation options we have used were rgb-to-gray, random-horizontal flip, random-adjust-brightness, random-adjust-contrast, random-crop-and-pad-image (min-area: 0.5, min-padded-size-ratio: [1,1], max-padded-size-ratio: [2,1]). Because random-crop-and-pad image augmentation option may result in a change of image size, we have provided batch size equal to 1 in training. Other parameters are kept unchanged from the values provided in the configuration file

3. https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md



(a) (b)

Figure 2: Confusion matrix for test set: (a) distribution in frames (b) normalized by the number images per class

that comes with the pre-trained model but we also provided them in the supplementary material. The same images provided to the classification network were used as input to the object detection network. The results of the experiments are provided in the next section.

3. Results

3.1. View Classification

The distribution of DICOM clips in our training, validation and test data belonging to each class are shown in Figure A. As can be seen there is a significant imbalance between classes. The exact numbers are 686 (Apical 2), 1061 (Apical 3), 2028 (Apical 4), 428 (Apical 5), 1416 (PLAX), 250 (PLAX-RVIF), 74 (PLAX-RVOT), 1882 (PSAX), 909 (PSAX-AoV), 451 (Subcostal). The overall accuracy we obtained was 97.62% on the test set. The confusion matrix obtained from the test set are shown in Figure 2. The diagonal elements show the number of correct predictions whereas off-diagonal elements show the number of misclassified images. The lowest accuracy we obtained was in the Apical 5 class, mostly because of high correlation to Apical 4 images. In addition, while the heart is contracting, the chamber appearing in the center (the aorta) can become very small in some frames, which makes the image look like an Apical 4 view. Also, we observed that a zoomed Apical 5 may look more like Apical 3. Next lowest accuracy values were obtained in the PLAX-RVIF and PLAX-RVOT views. Images in these classes exhibit large variations from patient to patient. Additional training data representing all these variations was yet not available.

3.2. Object Detection

In the evaluation of the object detection step, we have used mean average precision (mAP) and mean average recall (mAR) values calculated for the given intersection-over-union ratios (IoU). The evaluation results are given in Table Table 1. Here, mAP (IoU:0.50:0.95) corresponds to average mAP calculated over a range of IoU = 0.50:0.05:0.95. This metric is

MS COCO's standard detection metric. On the other hand, mAP (IoU:0.50) corresponds to mAP calculated at IoU=0.50 (PASCAL VOC's metric (Everingham et al.)). We have also shown results for mAP (IoU:0.75) and mAR (IoU:0.50:0.95). As can be seen from Table 1, the best results are obtained for Apical 4 and PLAX views because there are more samples available in training, and also the valves usually appear larger, which makes them easier to detect.

We also show some visual examples of bounding box detection applied to six different images selected from the test set per each cardiac view in Figure A. These images were hand-picked to represent different B-mode dynamic range, shifts, zoom factors, rotation and noise levels. In all of the figures corresponding to different views, the detection results of maximum score (> 0.5) and the prediction scores per class are shown. Apical 2, shown in the first row, consists of only MV. As can be seen, the valves were detected precisely in all the test images. The lowest score was 69% obtained in the second test image since there is more structure appearing around the valve, probably due to the imaging plane being close the wall of the heart. Apical 3, shown in the second row, consists of MV and AV. The scores for the AV are overall lower than the scores of MV because it is a smaller valve and sometimes hardly visible in especially noisy images such as in the last test image. Apical 4, shown in the third row, consist of MV and TV. When the heart is on a rotated plane, the MV appears smaller in size, which also decreases its detection probability with very low scores as seen in the third and fifth test images. In Apical 5, shown in the forth row, both MV and TV may appear partially but the region of clinical interest is left ventricle outflow tract (LVOT) location just above AV. Additionally, PLAX (fifth row) consists of MV and AV; and PLAX-RVIF (six row) consists of TV.For Apical 5, PLAX and PLAX-RVIF, the networks were able to precisely detect the objects with high scores.

Table 1: Evaluation results of object detection experiments

	Apical 2	Apical 3	Apical 4	Apical 5	PLAX	PLAX-RVIF
# test images	2164	2819	5303	1752	4030	1199
mAP (IoU:0.50:0.95)	0.151	0.170	0.343	0.166	0.463	0.217
mAP (IoU:0.50)	0.493	0.547	0.896	0.517	0.947	0.725
mAP (IoU:0.75)	0.041	0.042	0.146	0.058	0.365	0.048
mAR (IoU:0.50:0.95)	0.451	0.450	0.528	0.534	0.553	0.491

4. Conclusion

We have presented an end-to-end deep learning based pipeline that includes classification and object detection modules for the localization and identification of valves in cardiac ultrasound images. To the best of our knowledge, this is the first paper that uses state-of the art deep learning based object detection networks for this specific application. Our results suggest that it is possible to accurately locate and classify the valves using object detection techniques. For future work, we plan to add more training data with more precise annotations to increase the accuracy. We also plan to extend our work to enable it to work in more cardiac views.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE transactions on medical imaging*, 36(11):2204–2215, 2017.

Eva Costa, Nelson Martins, Malik Saad Sultan, Diana Veiga, Manuel Ferreira, Sandra Mattos, and Miguel Coimbra. Mitral valve leaflets segmentation in echocardiography using convolutional neural networks. In *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, pages 1–4. IEEE, 2019.

Bob D de Vos, Jelmer M Wolterink, Pim A de Jong, Tim Leiner, Max A Viergever, and Ivana Isgum. Convnet-based localization of anatomical structures in 3-d medical images. *IEEE Trans Med Imaging*, 36(7):1470–1481, 2017.

Yoni Dukler, Yurun Ge, Yizhou Qian, Shintaro Yamamoto, Baichuan Yuan, Long Zhao, Andrea L Bertozzi, Blake Hunter, Rafael Llerena, and Jesse T Yen. Automatic valve segmentation in cardiac ultrasound time series data. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741Y. International Society for Optics and Photonics, 2018.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Xiaohong Gao, Wei Li, Martin Loomes, and Lianyi Wang. A fused deep learning architecture for viewpoint classification of echocardiography. *Information Fusion*, 36:103–113, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, volume 4, 2017a.

Weilin Huang, Christopher P Bridge, J Alison Noble, and Andrew Zisserman. Temporal heartnet: towards human-level automatic analysis of fetal cardiac screening video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 341–349. Springer, 2017b.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick.

Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *npj Digital Medicine*, 1(1):6, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Vaanathi Sundaresan, Christopher P Bridge, Christos Ioannou, and J Alison Noble. Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 671–674. IEEE, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Ingmar Voigt, Mihai Scutaru, Tommaso Mansi, Bogdan Georgescu, Noha El-Zehiry, Helene Houle, and Dorin Comaniciu. Robust live tracking of mitral valve annulus for minimally-invasive intervention guidance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 439–446. Springer, 2015.

Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H Tison, Laura A Hallock, Lauren Beussink-Nelson, Mats H Lassen, Eugene Fan, Mandar A Aras, ChaRandle Jordan, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018.

Yefeng Zheng, Bogdan Georgescu, Haibin Ling, S Kevin Zhou, Michael Scheuering, and Dorin Comaniciu. Constrained marginal space learning for efficient 3d anatomical structure detection in medical images. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 194–201. IEEE, 2009.

Appendix A. Distribution of data and sample inference results

Figure 3: Distribution of DICOM clips in overall data

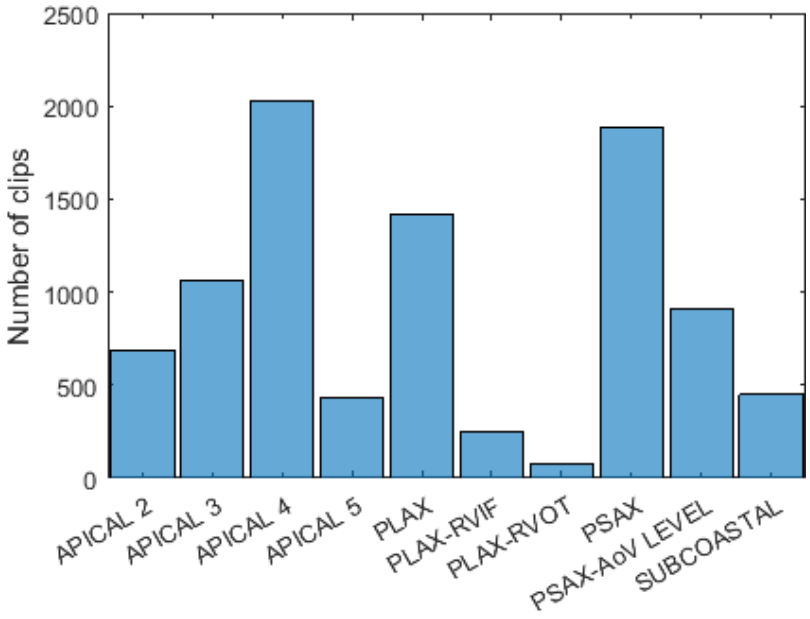


Figure 4: Predicted bounding boxes and prediction scores in percentage for Apical 2 (first row), Apical 3 (second row), Apical 4 (third row), Apical 5 (forth row), PLAX (fifth row) and PLAX-RVIF(sixth row) test images. MV, AV, TV and LVOT stand for Mitral Valve, Aortic Valve, Tricuspid Valve and Left Ventricular Outflow Tract, respectively.