# Domain and Geometry Agnostic CNNs for Left Atrium Segmentation in 3D Ultrasound

Markus A. Degel[1,2(✉)], Nassir Navab[1,3], and Shadi Albarqouni[1]

[1] Computer Aided Medical Procedures (CAMP), Technische Universität München, Munich, Germany
{markus.degel,shadi.albarqouni}@tum.de

[2] TOMTEC Imaging Systems GmbH, Unterschleissheim, Germany

[3] Whiting School of Engineering, Johns Hopkins University, Baltimore, USA

**Abstract.** Segmentation of the left atrium and deriving its size can help to predict and detect various cardiovascular conditions. Automation of this process in 3D Ultrasound image data is desirable, since manual delineations are time-consuming, challenging and observer-dependent. Convolutional neural networks have made improvements in computer vision and in medical image analysis. They have successfully been applied to segmentation tasks and were extended to work on volumetric data. In this paper we introduce a combined deep-learning based approach on volumetric segmentation in Ultrasound acquisitions with incorporation of prior knowledge about left atrial shape and imaging device. The results show, that including a shape prior helps the domain adaptation and the accuracy of segmentation is further increased with adversarial learning.

## 1 Introduction

Quantification of cardiac chambers and their functions stay the most important objective of cardiac imaging [7]. Left atrium (LA) physiology and function have an impact on the whole heart performance and its size is a valuable indicator for various cardiovascular conditions, such as atrial fibrillation (AF), stroke and diastolic dysfunction [7]. Compared to cardiac computed tomography (CCT) and cardiac magnetic resonance (CMR), as modalities to examine the heart, echocardiography provides wide availability, safety and good spatial and temporal resolution, without exposing the patients to harmful radiation. Volumetric measurements consider changes in all spatial dimensions, however, to obtain reproducible and accurate three-dimensional (3D) measurements, requires expert experience and is time consuming [4]. Automated segmentation and quantification could help to reduce inter/intra-observer variabilities and might also save costs and time in echocardiographic laboratories [4].

Previous automatic and semi-automatic approaches for LA segmentation have focused CCT and CMR as a planning and guidance tool for LA catheter interventions [1]. For 3D Ultrasound (US), the left ventricle (LV) was the

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11073, pp. 630–637, 2018.
https://doi.org/10.1007/978-3-030-00937-3_72

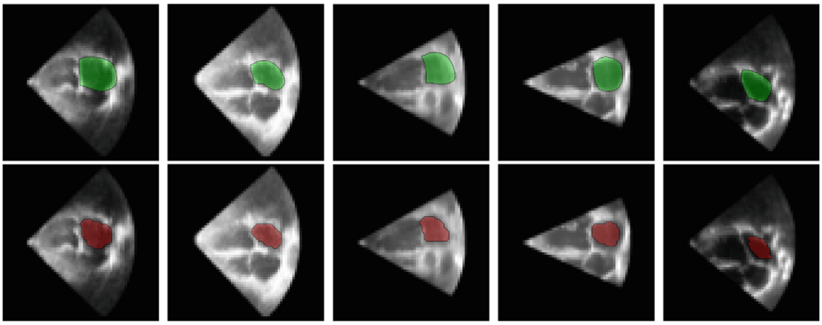

**Fig. 1.** Row 1: ground truth segmentation of LA, Row 2: prediction by *DGA* architecture (Table 3), Column 1 & 2: device EPIQ 7C (trained on Vivid E9), Column 3 & 4: device Vivid E9 (trained on EPIQ 7C), Column 5: device iE33 (trained on Vivid E9).

segmentation target, since its size and function remain the most important indication for a cardiac study [6]. LA segmentation in 3D US has not received much attention, apart from commercially available methods, which were also successfully validated against the gold standard CMR and CCT [3,10]. Almeida *et al.* [1] adapted a segmentation framework for LV, based on B-spline explicit active surfaces. Those methods, however, require more or less manual interaction. Recently, fully automatic segmentation of the left heart was validated against 2D and 3D echocardiography, as well as CCT [4].

Convolutional neural networks (CNN) and their special architectures of fully convolutional networks (FCN) have successfully been applied to the problem of medical image segmentation. Those networks are trained end-to-end, process the whole image and perform pixel-wise classification. The *V-Net* extends this idea to volumetric image data and enables 3D segmentation with the help of spatial convolutions, instead of processing the volumes slice-wise [8].

Automated segmentation in cardiac US images is challenging, due to artifacts caused by respiratory motion, shadows or signal-dropouts. Including shape priors in this task can help algorithms to yield more accurate and anatomically plausible results. Oktay *et al.* [9] introduced a way to incorporate such a prior with the help of an autoencoder network, that leads segmentation masks to follow an underlying shape representation.

Image data might be different (*e.g* with respect to resolution, contrast), due to varying imaging protocols and device manufacturers [2,5]. Although the segmentation task is equivalent, neural networks perform poorly when applied to data that was not available during training. Generating ground truth maps and retraining a new model for each domain is not a scalable solution. The problem of models to generalize to new image data can be approached by domain adaptation. Kamnitsas *et al.* [5] successfully introduced the application of unsupervised domain adaptation for different MRI databases, when an adversarial neural network was influencing the feature maps of a CNN, which was employed for a segmentation task.

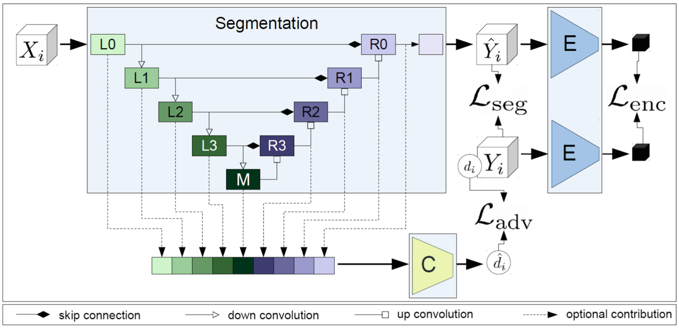

**Fig. 2.** Overview of the combined architecture: Image data $X_i$ is processed by *V-Net* [8]. $\mathcal{L}_{seg}$ is calculated from the resulting segmentation $\hat{Y}_i$ and the ground truth $Y_i$. Additionally, $\hat{Y}_i$ and $Y_i$ are encoded (E) to get the shape constraint. An optional number of feature maps, based on $X_i$ are extracted from *V-Net* to be processed in the classifier (C), which predicts a domain $\hat{d}_i$. Cross-entropy between $\hat{d}_i$ and the real domain $d_i$ determines the adversarial loss.

In this work, LA segmentation in 3D US volumes is performed with the help of neural networks. For the volumetric segmentation, *V-Net* will be trained, combined with additional losses, taking into account the geometrical constraint introduced by the shape of the LA and the desired ability to generalize to different US devices and settings.

## 2 Methodology

Our framework, as depicted in Fig. 2, consists of three existing methods; 3D Fully Convolutional Segmentation Network [8], Anatomic Constraint [9], and Domain Adaptation [5]. Nevertheless, it is a novelty to model the solution in a single framework, enabling analysis on the contribution of each element on the primary segmentation task. Further, the domain adaptation method has been leveraged to a 3D FCN segmentation framework, and applied successfully to the LA, showing a statistical significant improvement, as reported in Sect. 3.

***Segmentation.*** For the segmentation task, we employ *V-Net* [8] as a 3D FCN, which processes an image volume of size $n$, $X_i = \{x_1, ..., x_n\}$, $x_i \in \mathcal{X}$ and yields a segmentation mask $\hat{Y}_i = \{\hat{y}_1, ..., \hat{y}_n\}$, $\hat{y}_i \in \hat{\mathcal{Y}}$ in the original resolution. $\mathcal{X}$ represents the feature space of US acquisitions and $\hat{\mathcal{Y}}$ describes the probability of a voxel belonging to the segmentation.

The objective function of *V-Net* is adapted to the segmentation task. It is based on the Dice coefficient (Eq. 1), taking into account the possible imbalance of foreground to background, alleviating the need to re-weight samples.

$$\mathcal{L}_{seg} = 1 - \frac{2 \cdot \sum_i y_i \cdot \hat{y}_i}{\sum_i y_i^2 + \sum_i \hat{y}_i^2}, \tag{1}$$

with $\hat{y}_i$ being the prediction and $y_i$ the voxels of the ground truth $Y_i$ from the binary distribution $\mathcal{Y}$.

***Shape Prior.*** Incorporation of the shape prior to help the segmentation task is realized by training an autoencoder network on the segmentation ground truth masks $\boldsymbol{Y}$. The encoder reduces the label to a latent, low resolution representation $E(Y_i)$ and the decoder tries to retrieve the original volume $Y_i$. Due to the resolution reduction of the encoder, the shape information is encoded in a compact fashion [9].

During training, the output of the segmentation network $\hat{Y}_i$ is passed to the encoder, along with the ground truth label $Y_i$. Based on a distance metric $d(\cdot,\cdot)$, a loss between the latent codes of both inputs is calculated as

$$\mathcal{L}_{enc} = d(E(Y_i), E(\hat{Y}_i)). \tag{2}$$

The gradient is then back-propagated to the segmentation network.

***Domain Adaptation.*** When a network is trained on one type of data $\mathcal{X}_S$ (source domain) and evaluated on another $\mathcal{X}_T$ (target domain), the performance is poor in most cases. Domain invariant features are desired to make the segmentation network perform well on different data sets. Kamnitsas *et al.* [5] propose an approach to generate domain invariant features to increase a networks generalization capability.

Processing an image volume in a CNN yields a latent representation $h_l(X_i)$ after convolutional layer $l$. If the network is not domain invariant, those feature maps contain information about the data type (source or target domain). The idea to solve this issue, is to train a classifier $C$, which takes feature maps of the segmentation network as input and returns whether the input data was from source ($X_S$) or target ($X_T$) domain: $C(h_l(X_i)) = \hat{d}_i \in \{S, T\}$. The accuracy of this classifier is an indicator of how domain invariant the features are.

***Combination.*** The ideas introduced in the previous sections are combined to exploit the advantages of the individual approaches (Fig. 2). The loss of the domain classifier is used as an adversarial loss term, since the goal of the segmentation network is to lower the classification accuracy ($i.e$ maximize its loss). The inability of the classifier to tell, which type of data was segmented means that the feature maps are domain invariant. At the same time, $\mathcal{L}_{seg}$ and $\mathcal{L}_{enc}$ should be minimized. With $\mathcal{L}_{adv}$ as the binary cross entropy loss of the classifier $C$, this yields the following combined loss function:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda_{enc} \cdot \mathcal{L}_{enc} - \lambda_{adv} \cdot \mathcal{L}_{adv} \tag{3}$$

## 3 Experiments and Results

To evaluate the influence of different loss terms, we apply it to 3D Ultrasound data to perform end-systolic LA segmentation. The network is trained with images and labels from one device and tested on different devices.

**Table 1.** Data device and set distribution. iE33 datasets are only used for evaluation. Resolutions are equidistant. Resolution and opening angles of Ultrasound devices (azimuth & elevation) shown as: mean ± standard deviation.

| Property | EPIQ 7C | Vivid E9 | iE33 |
|---|---|---|---|
| Train/val/test | 33/7/27 | 39/8/32 | 0/0/15 |
| Resolution (mm/voxel) | $0.95 \pm 0.10$ | $0.95 \pm 0.10$ | $0.96 \pm 0.11$ |
| Azimuth (deg) | $87.1 \pm 4.7$ | $47.3 \pm 10.4$ | $80.2 \pm 0.0$ |
| Elevation (deg) | $78.2 \pm 0.1$ | $47.4 \pm 10.5$ | $91.6 \pm 0.0$ |

***Dataset.*** The data available for this work are 3D transthoracic echocardiography (TTE) examinations taken from clinical routine, which brings variations from differences in US imaging devices, protocols (resolution, opening angle) and patients (healthy, abnormal), raising the necessity of our proposed framework (Table 1). Multiple international centers contributed to a pool of 161 datasets, containing the LA ground truth segmentation in the entire recorded heart cycle, with the relevant phases for LA functionality (end-diastole, end-systole and pre-atrial contraction) identified.

Acquisition was performed with systems from GE (Vivid E9, GE Vingmed Ultrasound) and Philips (EPIQ 7C and iE33, Philips Medical Systems), each equipped with a matrix array transducer. Since there are only 15 datasets for device iE33, those examinations are not used for training, only for evaluation. The data is down-sampled, preserving angles and ratios, by zero padding (*cf.* Fig. 1), to enable processing of the entire volumes.

***Implementation.*** Network architectures are implemented using the TensorFlow[1] library (version 1.4) with GPU support. For our approach, the *V-Net* architecture is adapted, such that volumes of size 64 × 64 × 64 can be processed. The autoencoder network architecture is inspired from the one proposed in [9]. Feature maps from different levels and sizes are extracted from *V-Net* to be processed in the classifier (Fig. 2). By (repeated) application of convolutions of filter size 2 with stride 2, the feature maps are brought to the *V-Net* valley size (4 × 4 × 4), so they can be concatenated along the channel dimension.

***Training Details.*** The autoencoder network is trained before the combined training procedure, to obtain a meaningful latent representation for the shape prior. In the following training stages, the parameters of this network are frozen. The segmentation network is shortly pre-trained, as well as the classifier to introduce stability in the combined training and it can focus on realizing the scenario defined by the settings of $\lambda_{enc}$ and $\lambda_{adv}$. Feature maps L0, L2, M, R2 and R0 of the segmentation network are extracted for the classifier.

[1] https://www.tensorflow.org/.

**Table 2.** Training procedure details. Each training uses a learning rate decay of 0.99 after each epoch and a batch size of 4. $\boldsymbol{X} = \boldsymbol{X}_S \cup \boldsymbol{X}_T$, $\boldsymbol{d}$: domain labels.

| # | Name (parameters) | Optimizer | Learning rate | Weight reg. | Epochs | Data | Label |
|---|---|---|---|---|---|---|---|
| 1 | Autoencoder ($\theta_{ae}$) | Momentum $\beta$:0.9 | $5 \cdot 10^{-4}$ | 0.1 | 100 | $\boldsymbol{Y}_S$ | $\boldsymbol{Y}_S$ |
| 2 | Segmentation ($\theta_{seg}$) | Adam $\beta_1$: 0.99, $\beta_2$: 0.999 | $1 \cdot 10^{-5}$ | $5 \cdot 10^{-4}$ | 50 | $\boldsymbol{X}_S$ | $\boldsymbol{Y}_S$ |
| 3 | Classifier ($\theta_{adv}$) | SGD | $5 \cdot 10^{-5}$ | $1 \cdot 10^{-5}$ | 15 | $\boldsymbol{X}$ | $\boldsymbol{d}$ |
| 4 | Combination 3 ($\theta_{seg}$) | Momentum $\beta$:0.99 | $1 \cdot 10^{-5}$ | $5 \cdot 10^{-4}$ | 100 | $\boldsymbol{X}_S$ | $\boldsymbol{Y}_S$,$\boldsymbol{d}$ |
| | Classifier ($\theta_{adv}$) | SGD | $5 \cdot 10^{-5}$ | $1 \cdot 10^{-5}$ | | $\boldsymbol{X}$ | $\boldsymbol{d}$ |

The combined training procedure starts by adding $\mathcal{L}_{enc}$, for incorporation of the shape prior to the segmentation loss $\mathcal{L}_{seg}$. Adversarial influence begins after 10 epochs of combined training, linearly increasing $\lambda_{adv}$ until it reaches its maximum of 0.001 after another 10 epochs. While the combined training exclusively adjusts the parameters of the segmentation network $\theta_{seg}$, the classifier parameters $\theta_{adv}$ are continued to be trained in parallel to retain a potent adversarial loss term. A training overview is given in Table 2.

***Evaluation.*** The segmentation network returns a volume $\hat{Y}_i$ of probabilities for the voxels to belong to the foreground, *i.e* the segmentation of the LA. The threshold for the cutoff probability to obtain a binary segmentation mask is determined by the best Dice coefficient on the validation set, from which the biggest connected component is selected as the final LA segmentation.

Segmentation metrics [1,9] are reported in Table 3 for the recommended phase of LA segmentation (end-systole ES [7]). We refer to the *V-Net* architecture with the additional loss term $\mathcal{L}_{enc}$, calculated from the L2-distance ($d(p,q) = \|p - q\|_2^2$), as geometry agnostic CNN *GAL2*. To investigate the influence of a different distance metric, *GAACD* uses the angular cosine distance, as it was proposed in [2] (ACD, $d(p,q) = 1 - \frac{\sum_i p_i \cdot q_i}{\|p\|_2 \cdot \|q\|_2}$). Our domain and geometry agnostic CNN *DGA* leverages the better performing distance metric (ACD, based on test results) with the adversarial loss $\mathcal{L}_{adv}$. We define statistical significance based on the paired two-sample t-test on a 5% significance level.

When training on EPIQ 7C, *V-Net* performs better than the other architectures on the same device. However, those margins are not statistically significant (MSD: $p = 0.65$, HD: $p = 0.24$, DC: $P = 0.66$), compared to *DGA*. The increased performance of *DGA* compared to *V-Net* and *ACNN* is significant with respect to all metrics. Vivid E9 training yields *V-Net* with the best performance on the same device, with statistical significance on all metrics. *DGA* is significantly outperforming *V-Net* on EPIQ 7C in terms of MSD and HD. No significant differences are observable on the evaluation of device iE33. Independent of the distance metric utilized, an improvement in generalizability is observable compared to *V-Net* when the shape prior is included (*GAL2* & *GAACD*).

**Table 3.** Results for ES LA segmentation. Baseline *ACNN* and *V-Net* results are reported. *GAL2*: $\lambda_{adv} = 0$, $d$: L2-distance. *GAACD*: $\lambda_{adv} = 0$, $d$: ACD. *DGA*: $\lambda_{adv} = 0.001$, $d$: ACD. *GAL2*,*GAACD* & *DGA*: $\lambda_{enc} = 0.001$. Format: mean ± std.

| Training | Test | *V-Net* [8] | *ACNN*[9] | *GAL2* | *GAACD* | *DGA* |
|---|---|---|---|---|---|---|
| Mean Surface Distance (MSD) | | | | | | |
| EPIQ 7C | EPIQ 7C | **1.16±0.88** | 1.35 ± 1.19 | 1.26 ± 0.69 | 1.27 ± 0.69 | 1.21 ± 0.60 |
| | Vivid E9 | 3.56 ± 1.71 | 10.67 ± 7.29 | 3.87 ± 3.06 | 2.42 ± 1.32 | **2.01±1.63** |
| | iE33 | 1.44 ± 0.77 | **1.38±0.40** | 2.33 ± 2.38 | 1.94 ± 1.49 | 1.44 ± 0.35 |
| Vivid E9 | EPIQ 7C | 2.87 ± 1.53 | 4.39 ± 1.33 | 2.12 ± 0.96 | 1.87 ± 0.96 | **1.59±1.04** |
| | Vivid E9 | **0.94±0.59** | 1.57 ± 0.87 | 1.18 ± 0.38 | 1.12 ± 0.37 | 1.18 ± 0.37 |
| | iE33 | 4.72 ± 4.86 | 3.28 ± 2.22 | 4.18 ± 3.36 | 3.18 ± 2.88 | **2.62±1.46** |
| Hausdorff Distance (HD) | | | | | | |
| EPIQ 7C | EPIQ 7C | **4.46±2.73** | 5.52 ± 3.15 | 5.51 ± 2.31 | 5.33 ± 2.07 | 4.92 ± 1.60 |
| | Vivid E9 | 7.66 ± 2.94 | 16.87 ± 8.92 | 8.21 ± 5.06 | 5.79 ± 2.21 | **5.46±3.36** |
| | iE33 | **4.06±1.21** | 5.03 ± 1.39 | 5.60 ± 2.86 | 4.98 ± 2.02 | 4.70 ± 0.91 |
| Vivid E9 | EPIQ 7C | 10.82 ± 3.80 | 13.63 ± 2.87 | 8.09 ± 2.88 | 7.31 ± 2.51 | **5.47±2.45** |
| | Vivid E9 | **3.67±2.29** | 7.09 ± 3.21 | 5.41 ± 1.84 | 5.05 ± 1.70 | 5.14 ± 1.26 |
| | iE33 | 9.52 ± 6.44 | 11.60 ± 3.72 | 9.08 ± 3.64 | 7.13 ± 3.49 | **6.63±2.25** |
| Dice Coefficient (DC) | | | | | | |
| EPIQ 7C | EPIQ 7C | **0.75±0.17** | 0.69 ± 0.20 | 0.74 ± 0.10 | 0.73 ± 0.11 | 0.74 ± 0.10 |
| | Vivid E9 | 0.10 ± 0.21 | 0.15 ± 0.25 | 0.33 ± 0.27 | 0.32 ± 0.26 | **0.55±0.23** |
| | iE33 | 0.57 ± 0.31 | 0.64 ± 0.11 | 0.55 ± 0.19 | 0.59 ± 0.19 | **0.67±0.08** |
| Vivid E9 | EPIQ 7C | 0.56 ± 0.15 | 0.32 ± 0.18 | 0.59 ± 0.14 | 0.62 ± 0.17 | **0.63±0.17** |
| | Vivid E9 | **0.80±0.08** | 0.69 ± 0.11 | 0.73 ± 0.07 | 0.74 ± 0.08 | 0.73 ± 0.09 |
| | iE33 | 0.49 ± 0.37 | **0.50±0.16** | 0.38 ± 0.25 | 0.46 ± 0.27 | 0.46 ± 0.19 |

## 4 Discussion and Conclusion

While *V-Net* performs well on the task of LA segmentation, the ability to generalize to new domains is achieved by the introduction of a shape prior and the adversarial loss, as shown in the results. Including the shape prior boosts the segmentation performance on unseen devices and theoretically leads to a geometrically plausible segmentation in case of image artifacts. We ensure a potent classifier by training it in parallel to the *DGA* architecture. Thus, it can detect domain-specific features throughout the training procedure. The distance metric for the geometrical constraint is an interesting subject to further investigate, as well as extracting different *V-Net*-layers for processing in the classifier network.

**Acknowledgment.** We would like to thank Ozan Oktay for sharing his implementation of the *ACNN* architecture as a baseline of our project. Further, we thank Georg Schummers and Matthias Friedrichs from TOMTEC Imaging Systems GmbH for their support and helpful discussions.

## References

1. Almeida, N., et al.: Left-atrial segmentation from 3-D ultrasound using B-spline explicit active surfaces with scale uncoupling. IEEE UFFC-S **63**(2), 212–221 (2016)
2. Baur, C., Albarqouni, S., Navab, N.: Semi-supervised deep learning for fully convolutional networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 311–319. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_36
3. Buechel, R.R., et al.: Assessment of left atrial functional parameters using a novel dedicated analysis tool for real-time three-dimensional echocardiography: validation in comparison to magnetic resonance imaging. Int. J. Cardiovas Imag. **29**(3), 601–608 (2013)
4. van den Hoven, A.T., et al.: Transthoracic 3D echocardiographic left heart chamber quantification in patients with bicuspid aortic valve disease. Int. J. Cardiovas Imag. **33**(12), 1895–1903 (2017)
5. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47
6. Knackstedt, C., et al.: Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the FAST-EFs multicenter study. JACC **66**(13), 1456–1466 (2015)
7. Lang, R.M., et al.: Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. Eur. Heart J. **16**(3), 233–271 (2015)
8. Milletari, F., Navab, N., Ahmadi, S.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016)
9. Oktay, O., et al.: Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE TMI **37**(2), 384–395 (2018)
10. Rohner, A., et al.: Functional assessment of the left atrium by real-time three-dimensional echocardiography using a novel dedicated analysis tool: initial validation studies in comparison with computed tomography. Eur. J. Echocardiogr. **12**(7), 497–505 (2011)