Matwo-CapsNet: A Multi-Label Semantic Segmentation Capsules Network*

Savinien Bonheur[1,3,*], Darko Štern[2,3], Christian Payer[2,3], Michael Pienn[1], Horst Olschewski[1,4], and Martin Urschler[1,2,5]

[1] Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria
[2] Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria
[3] Institute of Computer Graphics and Vision, Graz University of Technology, Austria
[4] Department of Internal Medicine, Medical University of Graz, Austria
[5] School of Computer Science, University of Auckland, New Zealand

Abstract. Despite some design limitations, CNNs have been largely adopted by the computer vision community due to their efficacy and versatility. Introduced by Sabour et al. to circumvent some limitations of CNNs, capsules replace scalars with vectors to encode appearance feature representation, allowing better preservation of spatial relationships between whole objects and its parts. They also introduced the dynamic routing mechanism, which allows to weight the contributions of parts to a whole object differently at each inference step. Recently, Hinton et al. have proposed to solely encode pose information to model such part-whole relationships. Additionally, they used a matrix instead of a vector encoding in the capsules framework. In this work, we introduce several improvements to the capsules framework, allowing it to be applied for multi-label semantic segmentation. More specifically, we combine pose and appearance information encoded as matrices into a new type of capsule, i.e. Matwo-Caps. Additionally, we propose a novel routing mechanism, i.e. Dual Routing, which effectively combines these two kinds of information. We evaluate our resulting Matwo-CapsNet on the JSRT chest X-ray dataset by comparing it to SegCaps, a capsule based network for binary segmentation, as well as to other CNN based state-of-the-art segmentation methods, where we show that our Matwo-CapsNet achieves competitive results, while requiring only a fraction of the parameters of other previously proposed methods.

Keywords: capsules network, convolutional neural network, chest X-ray, multi-label, semantic segmentation

1 Introduction

Widely adopted by the computer vision and medical image analysis communities, convolutional neural networks (CNNs) have enabled huge progress in many applications related to these areas, e.g. image classification, computer aided diagnostics or semantic segmentation [5, 6]. However, CNNs also suffer from some

* Corresponding author: `savinien.bonheur@lvr.lbg.ac.at`

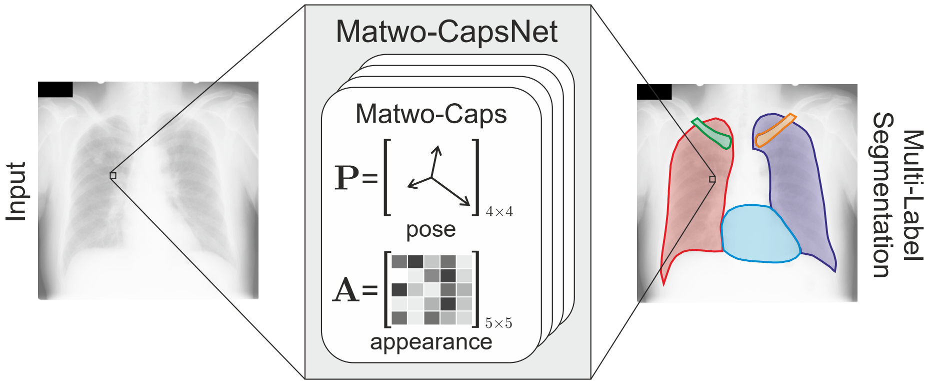


Fig. 1: Overview of our proposed multi-label semantic segmentation method, which incorporates Matwo-Caps consisting of a matrix **P** encoding pose information and a matrix **A** encoding appearance information.

limitations by design. Firstly, its formulation with consecutive convolution combined with pooling layers does not readily support the preservation of spatial dependencies of object parts in relation to the whole object. Especially in semantic segmentation applications, such dependencies may be crucial to encode constraints regarding anatomical information with the minimal amount of parameters, e.g. composing an X-ray image of the thorax by left and right lung structures as well as the heart, which are all anatomically constrained by each other in their relative locations. Secondly, CNNs max-pooling operation additionally leads not only to a loss of fine spatial information, but also potentially discards relevant information. Thirdly, the scalar representation of feature activations extracted with a CNN obscures its interpretability, which participates to the CNNs "black box" nature. Tackling these problems, Sabour et al. [10] proposed to replace scalar representations of feature activations with vectors encoding the feature instantiations, i.e. capsules [2]. Differently from CNNs, these capsules are coupled through dynamically calculated weights in each forward pass. This optimization mechanism, i.e. dynamic routing [10], allows to weight the contributions of parts to a whole object differently not just during training but also during inference. This interesting concept of capsule based networks has been mainly adapted for image classification applications, however, it has never been shown to be in line or go beyond state-of-the-art results. Moreover, up to now the capability of capsules being used for semantic segmentation has only been shown on binary segmentation tasks [4].

In this work we introduce Matwo-CapsNet, a multi-label semantic segmentation network that is based on the concept of capsules. Our proposed capsule network extends upon the related work as follows: Firstly, differently from the vector encoding of feature instantiations from [10], in our work we use a matrix encoding instead. Secondly, we combine this matrix encoding of feature instantiations from [10] with pose information inside each capsule. Although the use of the pose information encoded as a matrix has been proposed in [3], it has not yet been combined with feature information. Thirdly, to combine feature and pose information when learning spatial dependencies between capsules, we also propose a novel attention mechanism called Dual Routing. Finally, we extend the capsule network architecture from [4] to the multi-label segmentation task.

2 Method

Different to CNNs, where each object is represented as a scalar, capsule networks allow rich feature description by representing objects as vectors. Additionally, capsule networks are based on the hypothesis that a complex object (i.e. parent capsule) can be described through a weighted contribution of simpler objects (i.e. child capsules) after they are transformed into the feature space of the complex object. In the dynamic routing procedure, weights are dynamically calculated such that they correspond to the agreement of each transformed child capsule being a part of the complex parent capsule.

Thus, for each child-parent combination, a matrix $\mathbf{T}_{i \to n}$ of size $N \times N$ that transforms the child vector $\boldsymbol{v}_i$ of size N to the parent vector $\boldsymbol{v}_n$ of size N needs to be learned. Although in most of the previous works, capsules are represented as vectors, in [3] appearance encoding is replaced with a matrix $\mathbf{P}_i$ describing the object's pose.

Matwo-Capsule In this work, we combine these two concepts of representing an object by both appearance and pose, each having its own transformation matrix from child i to parent n. We encode as a matrix not just the pose $\mathbf{P}$, but also the appearance features $\mathbf{A}$, see Fig. 1. Thus, for the same number of transformation parameters ($N \times N$), we extend the representation of the appearance features $\mathbf{A}$ from a vector of size N to a matrix of size $N \times N$. The transformations of the child i pose $\mathbf{P}_i$ as well as appearance $\mathbf{A}_i$ matrices to the parent n matrices $\mathbf{P}_{i \to n}$ and $\mathbf{A}_{i \to n}$ are defined as

$$\mathbf{P}_{i \to n} = \mathbf{P}_i \mathbf{T}^{\mathbf{P}}_{i \to n} \quad \text{and} \quad \mathbf{A}_{i \to n} = \left(\mathbf{A}_i + b_{i \to n}\right) \mathbf{T}^{\mathbf{A}}_{i \to n}, \tag{1}$$

where $b_{i \to n}$ is a learned bias for the appearance matrix $\mathbf{A}_i$. $\mathbf{T}^{\mathbf{A}}_{i \to n}$ and $\mathbf{T}^{\mathbf{P}}_{i \to n}$ are the transformation matrices for appearance and pose, respectively. Same as the coordinate addition step in [3], we combine image coordinates x, y with each $\mathbf{T}^{\mathbf{P}}$.

In order to create the pose $\mathbf{P}_n$ and appearance matrix $\mathbf{A}_n$ of the parent capsule, the transformed matrices of all of its children need to be combined, i.e.,

$$\mathbf{P}_n = \mathit{Psquash}\left(\sum_i \alpha_{i \to n} \mathbf{P}_{i \to n}\right) \quad \text{and} \quad \mathbf{A}_n = \mathit{squash}\left(\sum_i \alpha_{i \to n} \mathbf{A}_{i \to n}\right), \tag{2}$$

where *Psquash* and *squash* are non-linear functions used to bound the values between -1 and 1. While we use the *squash* function as proposed in [10] for appearance matrices, we propose to use $\mathit{Psquash}(\mathbf{P}) = \frac{\mathbf{P}}{\max(\mathrm{abs}(\mathbf{P}))}$, a special

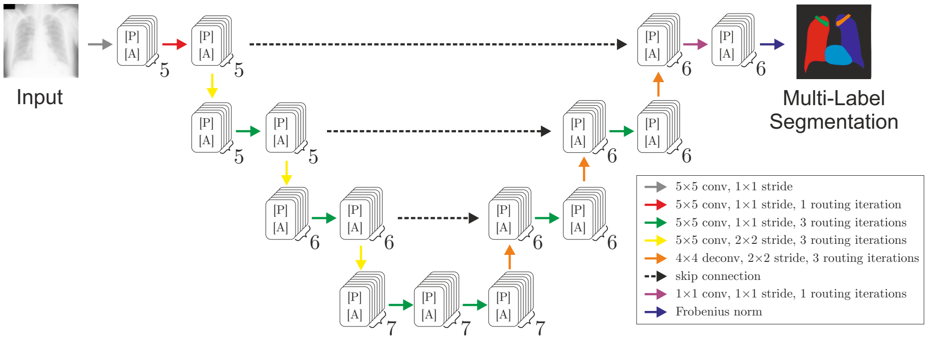

Input
[P]
[A]
5
6
7
Multi-Label
Segmentation
5×5 conv, 1×1 stride
5×5 conv, 1×1 stride, 1 routing iteration
5×5 conv, 1×1 stride, 3 routing iterations
5×5 conv, 2×2 stride, 3 routing iterations
4×4 deconv, 2×2 stride, 3 routing iterations
skip connection
1×1 conv, 1×1 stride, 1 routing iterations
Frobenius norm

Fig. 2: Our proposed Matwo-CapsNet architecture for multi-label segmentation. The number of capsules in each layer is indicated below the respective layer.

squashing function dedicated to pose matrices. The weighting factor $\alpha_{i\to n}$ of (2) defines how much each child contributes to the parent and is defined during the routing procedure.

Dual-Routing The dynamic routing mechanism of [10] follows an iterative optimization strategy based on the cross correlation between vectors of child and parent capsules to define their agreement $c_{i\to n}$. We extend this concept in our Dual Routing mechanism by treating the pose and appearance features separately before combining them via multiplication, i.e.,

$$c_{i\to n} = \langle \mathbf{P}_{i\to n}, \mathbf{P}_n \rangle_{\mathrm{F}} \cdot \langle \mathbf{A}_{i\to n}, \mathbf{A}_n \rangle_{\mathrm{F}}, \tag{3}$$

where $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ denotes the Frobenius inner product. Finally, the weighting factors $\alpha_{i\to n}$ for each child of equation (2) are calculated by applying the sigmoid function to $c_{i\to n}$.

Matwo-CapsNet Architecture We extend the SegCaps network architecture [4] for multi-label segmentation and integrate our Matwo-Caps into our proposed Matwo-CapsNet, see Fig. 2. Similarly as in SegCaps, for each pixel (x, y) of either the input image or the intermediate layers, a set of Matwo-Caps is defined, i.e. $\mathbf{P}_i(x, y)$ and $\mathbf{A}_i(x, y)$. To incorporate local neighborhood information inside of our Matwo-CapsNet, a convolution kernel of size $k \times k$ (see Fig. 2) is learned for the pose and appearance matrices $\mathbf{P}_i$ and $\mathbf{A}_i$. The predicted multi-label segmentation L at each location (x, y) corresponds to the index of the capsule of the last layer with the highest activation, i.e.,

$$L(x, y) = \arg\max_i (\|\mathbf{P}_i(x, y)\|_{\mathrm{F}} \cdot \|\mathbf{A}_i(x, y)\|_{\mathrm{F}}). \tag{4}$$

3 Experimental Setup and Results

Dataset We evaluate our Matwo-CapsNet on the Japanese Society of Radiological Technology (JSRT) dataset [11]. The JSRT dataset consists of 247 chest radiographs with a resolution of 2048×2048 and a pixel size of 0.175mm. The groundtruth segmentation labels were provided by van Ginneken et al. [1], who manually annotated left and right lungs, left and right clavicles and the heart, leading to six labels including the background. We split the JSRT dataset into two equally sized training and testing sets. Due to memory limitations, all images are scaled to a resolution of 128×128 pixels. As intensity preprocessing, the input images are rescaled such that the pixels of each image are within -1 and 1. To prevent overfitting, we apply random data augmentation in the form of spatial (translation, scaling, rotation, elastic deformations) and intensity (shift, scaling) transformations on the input images, as described in [8].

Evaluated Networks To ensure a fair comparison between capsule networks and CNNs, we also evaluate our implementation of the state-of-the-art *U-Net* [9] segmentation architecture. Differently to the original U-Net, we exchanged the deconvolution operations with linear upsampling, reduced the number of intermediate convolution outputs to 16, and reduced the number of levels to 4, leading to the approximate same number of parameters as our proposed Matwo-CapsNet. As a loss function, we use a pixel-wise softmax cross entropy.

We compare our proposed Matwo-CapsNet[1] with the *SegCaps* network [4], which was originally proposed for binary segmentation. We used the author's implementation from their source code repository[2] and reimplemented it in our network training framework to be consistent with our data augmentation. We also extend SegCaps to multi-label segmentation by increasing the number of output capsules from one to six. Furthermore, we increase the capacity of the multi-label SegCaps by having at least six capsules at any given layer, as well as increasing the length of the feature vector of each capsule to be at least $N = 32$. Different to [4], we do not incorporate a reconstruction loss and adapt SegCaps to the multi-label segmentation task by replacing the weighted binary cross-entropy loss with either a weighted softmax cross-entropy loss or a weighted spread loss.

To evaluate different contributions of our Matwo-CapsNet, we introduce two additional variants of the network, in which we replace the appearance matrix with a vector and use either dynamic routing as proposed by [10] (*MatVec-CapsNet* O_r) or our proposed Dual Routing (*MatVec-CapsNet* D_r). All our networks use the spread loss function.
Training our Matwo-CapsNet on the JSRT dataset with an NVidia Titan XP equipped with 12 GB RAM takes approximately 45 hours, while testing one image requires approximately one second.

[1] Our code is available at https://github.com/savinienb/Matwo-CapsNet

[2] https://github.com/lalonderodney/SegCaps

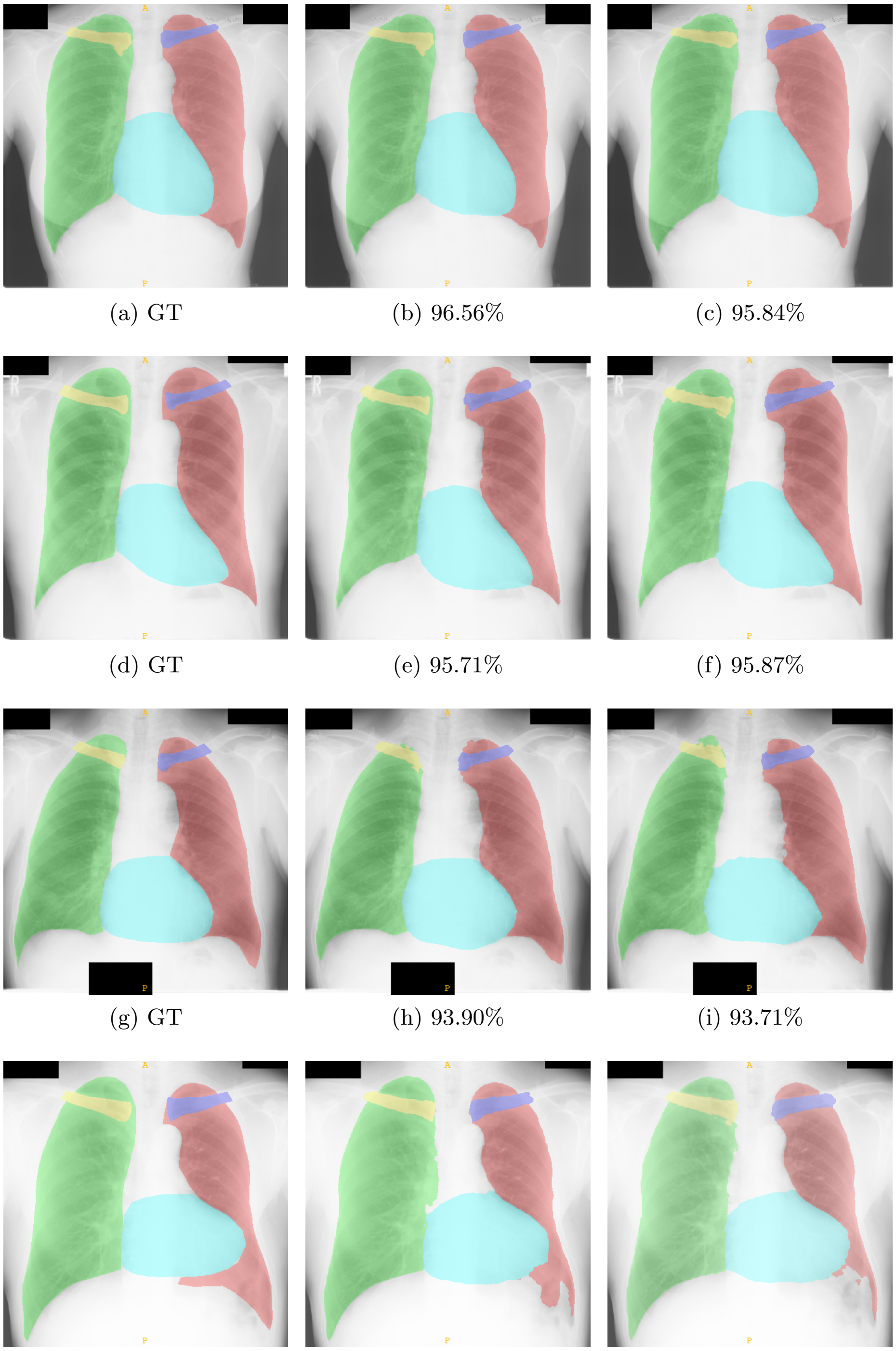
Table 1: The multi-label Dice scores of the evaluated networks in % on the JSRT dataset. The used loss function for each of our networks is shown within brackets. Number of network parameters are shown as multiples of thousands.

Network	#Params	Lungs		Clavicles		Heart
		L	R	L	R	
U-Net (Softmax)	42K	97.36	97.87	90.87	90.64	94.49
SegCaps (weighted Softmax)	2,129K	21.18	35.79	4.49	2.93	32.83
SegCaps (weighted Spread)	2,129K	30.74	0	0.06	0	23.23
MatVec-CapsNet O_r (Spread)	43K	95.57	96.43	82.89	82.56	92.37
MatVec-CapsNet D_r (Spread)	43K	96.60	97.15	86.41	86.38	93.42
Matwo-CapsNet (Spread)	43K	97.01	97.45	88.32	87.82	94.37
U-Net [7]	31,000K	96.4		83.4		93.4
InvertedNet [7]	3,141K	96.6		88.9		94.0

Results To verify that the original SegCaps-Net implementation is working within our augmentation and training framework, we evaluated SegCaps-Net on a binary task using the JSRT dataset, where the foreground object is defined as both left and right lungs and background as everything else. The results of this experiment show a Dice score of 95.38% for the foreground object. As outcome of our multi-label segmentation experiments, in Table 1 we show results in terms of multi-label Dice scores for the U-Net, our multi-label adaptations of the SegCaps-Net, and different variants of our proposed network, together with the state-of-the-art segmentation method for this dataset [7]. Note that different evaluation setups are used in [7]. Qualitative results are shown in Fig. 3, where the first row shows results where both U-Net and Matwo-CapsNet perform very well. The other rows show more challenging examples, where errors from the two methods are visualized.

4 Discussion and Conclusion

To the best of our knowledge, we are the first to show that multi-label semantic segmentation can be performed with a capsule based network architecture. The only other capsule based segmentation network proposed in the literature, i.e. SegCaps-Net [4], showed promising results but solely for binary lung segmentation when applied on a dataset of thoracic 2D CT slices. We tested their code within our framework on the JSRT dataset adapted for a binary segmentation task by setting left and right lung as foreground. This resulted in a Dice score of 95.38%, which is a competitive result when compared to state-of-the-art methods like the U-Net (see left and right lungs in Table 1). However, a direct extension of the SegCaps-Net to the multi-label segmentation task did not achieve satisfactory results, although we tested two different loss functions and compensated class label imbalances present in the JSRT dataset through the use of weighted loss functions.



(a) GT (b) 96.56% (c) 95.84%

(d) GT (e) 95.71% (f) 95.87%

(g) GT (h) 93.90% (i) 93.71%

(j) GT (k) 94.28% (l) 92.47%

Fig. 3: Example images for the JSRT dataset. Left images show the groundtruth labels (GT), center images show results and mean Dice scores of our U-Net implementation and right images show results and mean Dice scores of our proposed Matwo-SegCaps.

Using the same appearance vector encoding and dynamic routing mechanism as in [10], but by introducing a pose matrix and by extending the SegCaps architecture, we show how to successfully apply the capsule concept to multi-label segmentation (see *MatVec-CapsNet* O_r in Table 1). Simultaneously, the experiment shows that this performance is possible with a heavily reduced amount of network parameters as compared to SegCaps. Further, by replacing the routing mechanism used in [4, 10] with our proposed Dual Routing (*MatVec-CapsNet* D_r), we show that performance can be improved, especially for small anatomical structures, i.e. the clavicles. Finally, we receive our best results with the Matwo-CapsNet architecture, which additionally encodes the appearance information as a matrix instead of a vector. These results are very close to our heavily optimized U-Net implementation and both U-Net and Matwo-CapsNet outperform the currently best reported results on the JSRT dataset for images with the same resolution [7], while solely requiring a fraction of the network parameters. Our qualitative results presented in Fig. 3, show that both U-Net and Matwo-CapsNet have limitations with small structures like the right clavicle in (f) or the top of the right lung in (h) as well as with challenging pathological cases like the bottom of the left lung in (k) and (l).

In conclusion, our work has shown that representing appearance and pose information as matrix encodings, as well as combining both kinds of information using our novel Dual Routing mechanism, enables capsule based architectures to be used for multi-label segmentation. Moreover, we introduce a novel state-of-the-art U-Net architecture for multi-label segmentation of the JSRT dataset, which is highly optimized regarding its number of parameters. We compare our proposed capsule network with this architecture and demonstrate results that are in line for the multi-label segmentation task with a similar number of parameters. In future work, we will explore different, more complex routing schemes, e.g. EM-routing [3], and extend our Matwo-CapsNet to volumetric data.

References

1. van Ginneken, B., Stegmann, M.B., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Medical Image Analysis **10**(1), 19–40 (2006)
2. Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming Auto-Encoders. In: International Conference on Articial Neural Networks, pp. 44–51. Springer (2011)
3. Hinton, G.E., Sabour, S., Frosst, N.: Matrix capsules with EM routing. In: International Conference on Learning Representations (ICLR) (2018)
4. LaLonde, R., Bagci, U.: Capsules for Object Segmentation. In: International conference on Medical Imaging with Deep Learning (MIDL) (2018)
5. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
6. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
7. Novikov, A.A., Lenis, D., Major, D., Hladuvka, J., Wimmer, M., Bühler, K.: Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs. IEEE Transactions on Medical Imaging **37**(8), 1865–1876 (2018)

8. Payer, C., Štern, D., Bischof, H., Urschler, M.: Multi-Label Whole Heart Segmentation Using Anatomical Label Configurations. In: Pop, M. (ed.) Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. STACOM 2017. vol. 10663 LNCS, pp. 190–198. Springer, Cham (2018)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Proceedings Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241 (2015)
10. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic Routing Between Capsules. In: Neural Information Processing Systems (NIPS) (2017)
11. Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K.: Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule. American Journal of Roentgenology **174**(1), 71–74 (2000)