Efficient multi-modal whole heart segmentation via cascaded U-Net: a practical solution for clinical settings

Raul J. Salgado-Garcia[1], Nicolás Vila-Blanco[1,2], María J. Carreira[1,2], and Marta Nuñez-Garcia[1,2]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782, Santiago de Compostela, Spain
[2] Departamento de Electrónica e Computación, Escuela Técnica Superior de Ingeniería, Universidade de Santiago de Compostela, Spain

Abstract. Cardiovascular analysis based on cardiac imaging highly benefits from automated segmentation methods. The use of deep learning is currently the most effective approach. However, its utilization in medical settings is frequently constrained by the unavailability of high-capacity hardware resources, and the high variability of medical images, which challenges the generalization ability of deep learning techniques. We propose a pipeline of two sequential U-Net for CT and MRI segmentation, configured with low complexity, allowing for usability in clinical practice. In the first stage, single-label segmentation is used to crop the image volume to a bounding box surrounding the heart. The second stage focuses on the detected region of interest. Multi-label segmentation is performed on the trimmed volume to extract 7 different substructures of the heart. Results from the WHS++ challenge validation phase show that our method achieves an average Dice Similarity Coefficient of 0.9311 on CT, and 0.8652 on MRI data. Importantly, the inference times are kept to a minimum, even when using CPU computing (∼7 s).

Keywords: Cardiac segmentation · Convolutional neural network · Whole heart segmentation · CT · MRI.

1 Introduction

Cardiovascular diseases are a major global concern. They are the primary cause of mortality globally, with 19.4 million deaths in 2021, having a significant impact on various domains, including health, society, politics, the environment, and the economy [3, 12]. Different approaches and technologies have been proposed for improving cardiovascular disease diagnosis and treatment, including computational methods for automated 3D cardiac image analysis.

Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) images provide highly valuable anatomic information that needs to be properly extracted, processed and interpreted for clinical practice and research. Many

segmentation methods in the literature have focused on single cardiac substructures, such as the left ventricle [10] or the left atrium [13], due to their relation to a particular pathology, such as myocardial infarction [9] or atrial fibrillation [8]. In contrast, whole heart segmentation (WHS) involves the extraction of several main cardiac substructures, thereby enabling the investigation of a wide range of cardiac pathology.

The U-Net[7], a convolutional neural network, has become a de facto standard for many medical segmentation tasks [11, 1, 6]. However, the segmentation of the heart is particularly challenging, as compared to other organs, due to the shape variability observed among patients (attributable to normal human variability, disease, aging, treatments, etc.) and during the cardiac cycle; and poor image quality due to noise and image artifacts. Furthermore, the different modalities of CT and MRI, as well as domain shifts in multi-center data, present a challenge to the generalization ability of the various segmentation approaches.

Also, the high dimensionality of 3D images, especially on CT, makes it impossible to treat the whole image simultaneously due to memory and time limitations. As a consequence, data subsampling is required, which impacts on the quality of the resulting segmentations. An alternative that allows the use of higher resolutions without high memory consumption was proposed in [5]. The authors used a pipeline of two stages where in the first one the image is cropped to the region of interest so that the final segmentation can be performed in higher resolution due to prior elimination of the useless information. Similarly, the cascade configuration of the well-known self-configurable nnU-Net [4] utilizes first a 3D U-Net on low resolution images, and then a second high-resolution 3D U-Net that refines the segmentations. While high-complexity solutions are typically accurate (e.g., the nn-UNet and extensions) they may not be a practical solution in clinical settings where hardware resources are often limited.

The WHS++ track[3] of the Comprehensive Analysis & computing of REal-world medical images challenge (CARE2024)[4], held in conjunction with MICCAI 2024[5], was proposed to objectively compare novel approaches for whole heart segmentation [16, 15, 14, 2]. In particular, 7 cardiac substructures were considered: the left and right ventricular blood cavities, the left and right atrial blood cavities, the myocardium of the left ventricle, the ascending aorta, and the pulmonary artery. The dataset comprises 206 multi-modal (CT and MRI) whole heart volumes from 6 different centers and covering different cardiac diseases.

In the context of this competition, we propose to use state-of-the-art methods that can be implemented in clinical facilities. The aim of this decision is to check if more complex options are worth the increased computational cost. We propose a U-Net-based two-stage pipeline. First, all heart substructures are segmented as a whole, and the bounding box surrounding this segmentation is used to crop the volumetric image. Then, the cropped image is segmented to isolate the 7 heart substructures. We investigated different network configurations to find the

[3] http://www.zmic.org.cn/care_2024/track5/
[4] http://www.zmic.org.cn/care_2024/
[5] https://conferences.miccai.org/2024/en/

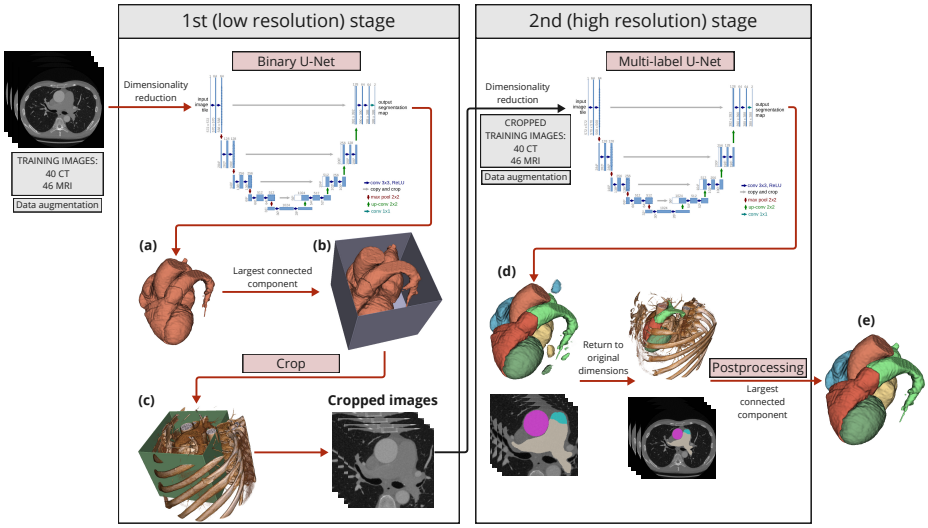

Fig. 1: Overall pipeline. In the low-resolution stage (left), the heart is segmented as a whole, and the image is cropped to the bounding box surrounding this segmentation. Then, in the high-resolution stage (right), the cropped image is used to obtain the final multi-label segmentation. U-Net scheme from [7].

best compromise between high accuracy and low complexity, and compared the results with a single-stage (only 1 U-Net) approach.

2 Methodology

The proposed pipeline can be seen in Fig. 1. Similarly to Jia et al. [5], we first use a 3D U-Net to locate the region of interest, which is a box containing the heart. The labels of the 7 substructures are combined to create a single label mask of the entire heart. Input image intensities are normalized, and images and ground truth segmentations are resized to $128 \times 128 \times 128$, keeping memory consumption low but limiting the loss of spatial information. The output of the U-Net is a single label segmentation of the heart, from which only the largest connected component is kept (Fig. 1-a). A bounding box surrounding this segmentation is computed (Fig. 1-b) and used to crop the original (i.e., not resized) volumetric image (Fig. 1-c). For cropping the image, we investigated 3 different approaches: a fixed crop for all images; a variable crop adjusted to each heart and; a variable crop adding a margin of 15% of the initial bounding box size.

In the second (high resolution) stage, another 3D U-Net is used, but in a different manner. It performs whole heart segmentation by taking the cropped image and assigning to each substructure a different label (Fig. 1-d). Importantly, by cropping the image and retaining solely the region of interest, it is possible to utilize a higher resolution without an increase in memory consumption. The

output of the network is a multi-label segmentation, and only the largest 3D connected component of each substructure is kept, avoiding separated unrealistic particles (Fig. 1-e).

For training both U-Nets, we use data augmentation (noise, scaling, rotation and elastic deformations), the Adam optimizer, the cross-entropy loss function, and a mini-batch size set to 1 to work with the highest possible resolution.

3 Experiments and Results

The method was evaluated on the WHS++ challenge dataset. It comprises 206 multi-modal (104 CT, and 102 MRI) whole heart volumes that were split by the organizers as follows: training set (40 CT and 46 MRI), validation set (30 CT and 20 MRI), and test set (34 CT and 36 MRI). Participants were asked to evaluate the quality of their results using the challenge's platform where different metrics were used: the Dice Similarity Coefficient (DSC), to evaluate the degree of overlap between the estimated and ground truth segmentations; and the Average Symmetric Surface Distance (ASSD) and Hausdorff Distance (HD) to ascertain the quality of the boundaries. ASSD represents the average distance between the segmentation and the ground truth, with a smaller ASSD indicating higher segmentation accuracy. In contrast, HD quantifies the longest distance, thereby reflecting the maximum segmentation error committed.

Training phase. We tested different U-Net configurations, such as the number of starting filters and depth, and we performed six-fold cross-validation. Differences across architectures were tested using paired t-tests.

For the first (low resolution) stage U-Net we tested depth levels of 4 and 5; and 32 and 64 start filters. As the objective of this first stage is not to perform a perfect segmentation but to yield an accurate heart crop, the different configurations were compared by means of the Intersection over Union (IoU) of the real and predicted bounding boxes. We found that there were not statistically significant differences between them (at the 5% level of significance (i.e., $p > 0.05$)) and we chose the least complex solution: 4 depth and 32 start filters. Importantly, there is no need for very accurate results in this first stage, as small errors will not significantly change the computed bounding box. Nevertheless, we computed performance metrics at this stage and this U-Net achieved a mean IoU of 0.8774 ± 0.0967 on CT, and 0.7214 ± 0.1129 on MRI data.

With respect to the cropping of the volumetric image, our experiments showed that a fixed crop, using the size of the biggest heart in the dataset, often includes a lot of information that is not of interest (Fig. 2-a). On the contrary, a variable crop using every individual's heart size removes all unnecessary information, and allows for the use of higher resolution in the second stage (Fig. 2-b). However, slight errors in the prior low resolution segmentation may lead to loss of information (e.g., incomplete heart). To overcome this issue, a margin of 15% of the box size was added to the bounding box obtained using the variable crop (Fig. 2-c). This solution represents a good balance between irrelevant information inclusion

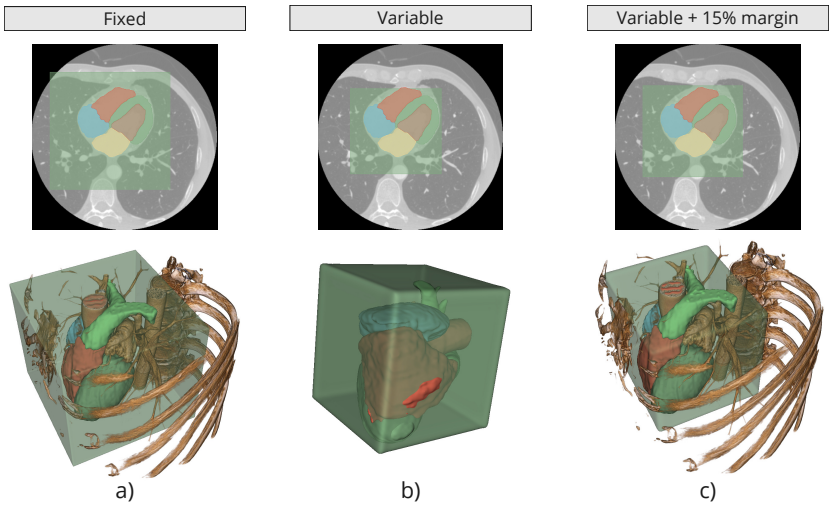

Fig. 2: Depiction of the 3 suggested crops in an example case: a) fixed crop where extra irrelevant information is included; b) variable crop where it can be noted the loss of information on the right ventricle (red label); and c) variable crop with 15% margin representing a good balance between completeness and irrelevant information exclusion. The computed bounding boxes are shown superimposed in green. Bones and other structures are shown for spatial reference.

and data completeness, promotes the reasonable use of computational resources, and it was therefore the option included in our framework.

In the second (high resolution) stage U-Net, we tested depths of 4, 5 and 6, and 32, 64 and 128 start filters. We downsampled images to isotropic sizes of 128, 192 and 256. The best configuration in cross-validation was depth level of 5, 64 start filters, and image size equal to $192 \times 192 \times 192$ (mean DSC $= 0.8561 \pm 0.0730$ on CT, and 0.8082 ± 0.0590 on MRI). To obtain the final segmentation model, we trained the U-Net with this configuration on a fixed number of 110 epochs using the full training set. In addition, since the evaluation platform allowed for multiple submissions, we also trained models using different network configurations (the same as in cross-validation).

Validation phase. The trained models were used for predicting the segmentations of the validation set, that were then uploaded to the evaluation platform of the challenge. DSC results for each heart substructure under consideration, on MRI and CT data, are shown in Table 1. The model demonstrated good accuracy in all substructures for both CT and MRI data, with low variability across cases (except for some structures in MRI) showing its reliability and robustness. Results were, however, more accurate for CT images, especially for the ascending aorta and the left ventricle. For MRI images, the highest quality segmentations were achieved for the right ventricle, the right atrium, and the ascending aorta.

Table 1: Best DSC results (mean ± standard deviation) in the validation phase corresponding to the 7 heart substructures separately. LV/RV = Left/Right Ventricle; Myo = Myocardiun; LA/RA = Left/Right Atrium; AO = Ascending Aorta; PA = Pulmonary Artery.

Substructure	CT	MRI
LV	0.9455 ± 0.0296	0.8723 ± 0.2077
Myo	0.9275 ± 0.0301	0.8082 ± 0.1074
RV	0.9139 ± 0.0353	0.8890 ± 0.0652
LA	0.9384 ± 0.0360	0.8428 ± 0.2004
RA	0.9145 ± 0.0432	0.8872 ± 0.0454
AO	0.9562 ± 0.0187	0.8805 ± 0.0598
PA	0.8598 ± 0.0693	0.7859 ± 0.1781

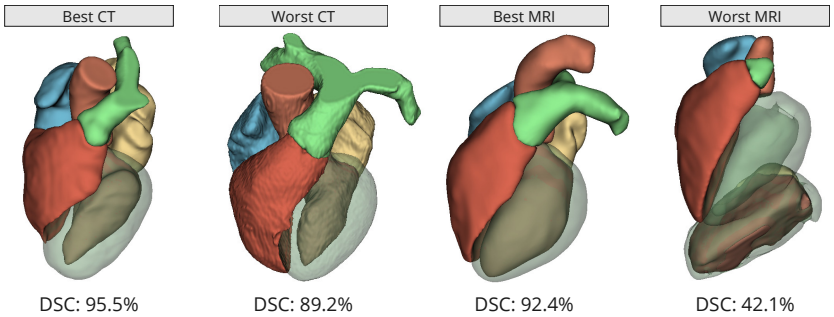

Fig. 3: 3D models of the best and worst segmentation results for CT (2 leftmost), and MRI (2 rightmost) on the validation set, according to the DSC metric.

On the contrary, the difficulty of pulmonary artery segmentation can be inferred by the lowest accuracy obtained in the two modalities. Fig. 3 depicts segmentation results corresponding to the best and worst CT and MRI validation cases. It can be observed that even the worst CT result in this dataset may be useful in clinical practice, while some MRI results are poor (Fig. 3 rightmost).

WHS results (i.e., the average of the 7 substructures) are shown in Table 2 (several tests corresponding to different configuration of the 2nd stage U-Net). It can be seen that, as found in the training phase, the best configuration (first row) had a depth of 5 levels, 64 start filters, and an image size of 192 × 192 × 192. Next we describe the quantitative results (mean ± standard deviation), along with the corresponding position in the ranking (#) at the time of this paper's submission, considering only one submission per team (the best one according to the leader board). As shown in Table 2, with respect to DSC, our methodology is accurate, especially for CT data (CT: 0.9311 ± 0.0182 (#7); MRI: 0.8652 ± 0.1075 (#7)), and it outperforms a single-stage (one U-Net) approach (CT: 0.9181 ± 0.0240; MRI: 0.8496 ± 0.0545). Fig. 4 depicts a boxplot representation of DSC on the validation set, showing that there are only 2 poor

Table 2: Results of our 4 best submissions, sorted from best (on top) to worse. The U-Net configuration (2nd stage) is specified as: depth-number start filters-dimensions. NP = No Postprocessing; DSC = Dice Score; ASSD = Average Symmetric Surface Distance; HD = Hausdorff Distance.

Config.	CT DSC	CT HD	CT ASSD	MRI DSC	MRI HD	MRI ASSD
5-64-192	0.9311	12.7226	0.8569	0.8652	21.4843	1.8424
6-64-192	0.9285	14.159	0.8638	0.8693	20.0673	1.7204
4-64-192	0.9283	13.7645	0.8814	0.8603	22.6692	1.9127
4-64-192 NP	0.9279	14.4748	0.8838	0.8668	21.7604	1.6088
single-stage	0.9181	13.5829	0.9608	0.8496	19.8720	1.6852

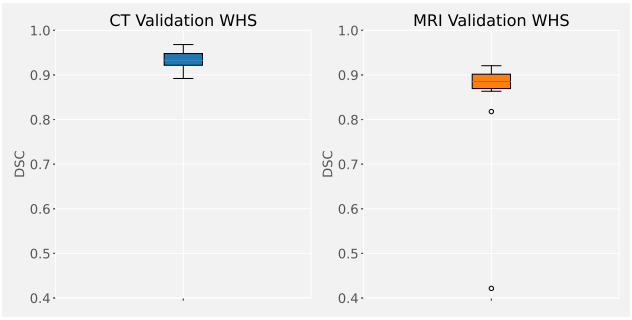

Fig. 4: Boxplot representation of DSC on the validation set on CT (left) and MRI (right) data.

results, depicted as outliers. With respect to HD, the results were remarkable for CT (12.7226 ± 4.8852 mm) ranking first in the leader board. For MRI data, the values were higher (i.e., worse) (21.4885 ± 11.1319 mm (#6)). With respect to ASSD, results were also better (i.e., lower) for CT data (CT: 0.8569 ± 0.2321 mm (#8); MRI: 1.8424 ± 2.4822 mm (#6)) while the performance was better in MRI than in CT as compared to the other teams. Interestingly, for MRI data and mainly according to ASSD and HD, both the no postprocessing and single-stage approaches outperformed some of the other options. This result may be related to the impact of the two outliers shown in Fig. 4 in the small MRI validation set (20 instances). Overall, combining all the metrics and according to the WHS++ evaluation platform, our team ranked 6th out of 12. Results corresponding to the

Table 3: Leader board showing the results of the best two teams and ours. The best result of each metric is shown in bold.

#	CT DSC	CT HD	CT ASSD	MRI DSC	MRI HD	MRI ASSD
1	**0.9439**	12.9392	**0.6945**	**0.8959**	17.283	**1.2083**
2	0.9333	13.9054	0.8388	0.893	**16.1008**	1.2411
6 (Ours)	0.9311	**12.7226**	0.8569	0.8652	21.4843	1.8424

best two participants and our best result can be seen in Table 3. We consider it would be beneficial to perform statistical significance tests to ascertain whether the discrepancy between the various tests is statistically significant or if they are equally valid solutions.

The time efficiency of our method is evidenced by the inference times of the networks: the first (low resolution single label) U-Net lasted on average 0.0026 ± 0.0000 s and 0.7339 ± 0.0508 s when using GPU (NVIDIA A100) or CPU (Intel Xeon Ice Lake 8352Y) computing, respectively; and the second (high resolution multi-label) U-Net 0.0032 ± 0.0000 s (GPU) and 6.5399 ± 0.0666 s (CPU). The run times of the other operations were negligible.

4 Discussion

We have developed a WHS method pursuing maximum efficiency while ensuring high-quality and reliable results. Accordingly, we chose a simple configuration for the first U-Net, and added a margin to the computed bounding box to reduce the impact of the first network's accuracy on the resulting segmentations. The second U-Net is tasked with achieving the highest possible quality in the final segmentation, and it is therefore more complex than the first one.

Results on the validation set show that our approach can make accurate predictions, specially on CT data. The performance difference between CT and MRI is especially notable in the myocardium, maybe because in MRI the interface between the LV blood pool and the myocardial wall is harder to differentiate. Also, MRI images typically contain more acquisition artifacts and have higher variability which makes accurate segmentation more challenging. Finally, CTA images (one of the CT modalities in the dataset) use an injection of contrast to provide better visualization of the vessels, which facilitates their segmentation.

The proposed method allows for rapid acquisition of results, thus enabling its utilization in medical settings. As compared to the single-stage approach, the simplicity of the first U-Net and the cropping strategy adds minimal complexity while achieving better results. By avoiding the use of ensembles or patch-wise segmentation, our method involves only 2 models of 1.2 GB in total and 2 inferences (1 at each stage). This results in minimal inference times when high-capacity hardware is available ($\sim$1 s), and also when a single CPU, accessible in any hospital, is utilized ($\sim$7 s). Preliminary tests show that this is much more efficient than the 10 models used in the nnU-Net, that total 2.4 GB and perform more than 100 inferences, even on the smallest images.

In this work, we have evaluated the impact of a few parameters, but other aspects, such as explicitly considering the multi-modality of the data or using optimized loss functions, are worth exploring.

5 Conclusions

In this paper, we have shown that a two-stage U-Net cascade pipeline for WHS provides accurate results while keeping inference times minimal. The configura-

tion was kept as simple as possible, demonstrating that specifically addressing data multi-modality or using more complex configurations, architectures, or loss functions may not be essential for achieving a segmentation model usable in clinical practice. The proposed methodology has obtained fast and accurate results in the validation phase of the WHS++ challenge, proving its usability even when no high-capacity hardware is available.

Acknowledgments. This work has received financial support from the Galician Ministry of Culture, Education, Professional Training and University (Galician Research Center accreditation 2024-2027 ED431G2023/04; and Reference Competitive Group accreditations ED431C2021/48 and ED431C2022/19). The previous grants are co-funded by the European Regional Development Fund (ERDF/FEDER program). This work was also supported by the Spanish Ministry of Science and Innovation, the Spanish Research Agency, and the European Union through the "Ramón y Cajal" 2022 program (RYC2022-035469-I).

References

1. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of U-Net. arXiv preprint arXiv:2211.14830 (2022)
2. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. Medical Image Analysis **89**, 102889 (2023)
3. Institute for Health Metrics and Evaluation: Global health metrics. cardiovascular diseases - level 2 cause. `https://www.healthdata.org/research-analysis/diseases-injuries-risks/factsheets/2021-cardiovascular-diseases-level-2-disease`
4. Isensee, F., Jaeger, P., Kohl, S., Petersen, J., Maier-Hein, K.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
5. Jia, S., Despinasse, A., Wang, Z., Delingette, H., Pennec, X., Jaïs, P., Cochet, H., Sermesant, M.: Automatically segmenting the left atrium from cardiac images using successive 3D U-nets and a contour loss. In: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018. pp. 221–229 (2019)
6. Komosar, A., Stefanović, D., Sladojević, S.: An overview of image processing in biomedicine using U-Net convolutional neural network architecture. Journal of Computer and Forensic Sciences (2024)
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241 (2015)
8. Sagris, M., Vardas, E.P., Theofilis, P., Antonopoulos, A.S., Oikonomou, E., Tousoulis, D.: Atrial fibrillation: pathogenesis, predisposing factors, and genetics. International journal of molecular sciences **23**(1), 6 (2021)
9. Salari, N., Morddarvanjoghi, F., Abdolmaleki, A., Rasoulpoor, S., Khaleghi, A.A., Hezarkhani, L.A., Shohaimi, S., Mohammadi, M.: The global prevalence of myocardial infarction: a systematic review and meta-analysis. BMC cardiovascular disorders **23**(1), 206 (2023)

10. Shoaib, M.A., Chuah, J.H., Ali, R., Hasikin, K., Khalil, A., Hum, Y.C., Tee, Y.K., Dhanalakshmi, S., Lai, K.W.: An overview of deep learning methods for left ventricle segmentation. Computational intelligence and neuroscience **2023**(1), 4208231 (2023)
11. Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V.: U-net and its variants for medical image segmentation: A review of theory and applications. IEEE access **9**, 82031–82057 (2021)
12. Vaduganathan, M., Mensah, G.A., Turco, J.V., Fuster, V., Roth, G.A.: The global burden of cardiovascular diseases and risk: a compass for future health (2022)
13. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., Heng, P.A., Ni, D., Li, C., Tong, Q., Si, W., Puybareau, E., Khoudli, Y., Géraud, T., Chen, C., Bai, W., Rueckert, D., Xu, L., Zhuang, X., Luo, X., Jia, S., Sermesant, M., Liu, Y., Wang, K., Borra, D., Masci, A., Corsi, C., de Vente, C., Veta, M., Karim, R., Preetha, C., Engelhardt, S., Qiao, M., Wang, Y., Tao, Q., Nuñez-Garcia, M., Camara, O., Savioli, N., Lamata, P., Zhao, J.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical image analysis **67**, 101832 (2021)
14. Zhuang, X.: Multivariate Mixture Model for Myocardial Segmentation Combining Multi-Source Images. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(12), 2933–2946 (2019)
15. Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Medical image analysis **58**, 101537 (2019)
16. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Medical Image Analysis **31**, 77–87 (2016)