ImageCHD: A 3D Computed Tomography Image Dataset for Classification of Congenital Heart Disease

Xiaowei Xu[1], Tianchen Wang[2], Jian Zhuang[1], Haiyun Yuan[1], Meiping Huang[1], Jianzheng Cen[1], Qianjun Jia[1], Yuhao Dong[1], and Yiyu Shi[2(✉)]

[1] Guangdong Provincial People's Hospital, Guangzhou, China
`xiao.wei.xu@foxmail.com`, `zhuangjian5413@tom.com`, `huangmeiping@126.com`
[2] University of Notre Dame, Notre Dame, USA
`{twang9,yshi4}@nd.edu`

Abstract. Congenital heart disease (CHD) is the most common type of birth defects, which occurs 1 in every 110 births in the United States. CHD usually comes with severe variations in heart structure and great artery connections that can be classified into many types. Thus highly specialized domain knowledge and time-consuming human process is needed to analyze the associated medical images. On the other hand, due to the complexity of CHD and the lack of dataset, little has been explored on the automatic diagnosis (classification) of CHDs. In this paper, we present ImageCHD, the first medical image dataset for CHD classification. ImageCHD contains 110 3D Computed Tomography (CT) images covering most types of CHD, which is of decent size compared with existing medical imaging datasets. Classification of CHDs requires the identification of large structural changes without any local tissue changes, with limited data. It is an example of a larger class of problems that are quite difficult for current machine-learning based vision methods to solve. To demonstrate this, we further present a baseline framework for automatic classification of CHD, based on a state-of-the-art CHD segmentation method. Experimental results show that the baseline framework can only achieve a classification accuracy of 82.0% under selective prediction scheme with 88.4% coverage, leaving big room for further improvement. We hope that ImageCHD can stimulate further research and lead to innovative and generic solutions that would have an impact in multiple domains. Our dataset is released to the public [1].

Keywords: Dataset · Congenital heart disease · Automatic diagnosis · Computed tomography

1 Introduction

Congenital heart disease (CHD) is the problem with the heart structure that is present at birth, which is the most common type of birth defects [3]. In recent

years, noninvasive imaging techniques such as computed tomography (CT) have prevailed in comprehensive diagnosis, intervention decision-making, and regular follow-up for CHD. However, analysis (e.g., segmentation or classification) of these medical images are usually performed manually by experienced cardiovascular radiologists, which is time-consuming and requires highly specialized domain knowledge.
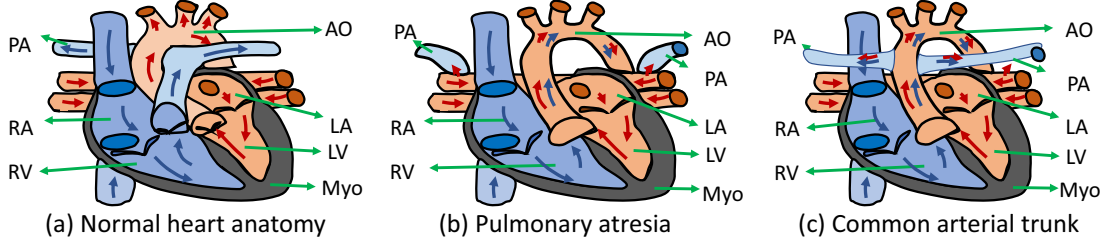

(a) Normal heart anatomy   (b) Pulmonary atresia   (c) Common arterial trunk

Fig. 1. Examples of large heart structure and great artery connection variations in CHD (LV-left ventricle, RV-right ventricle, LA-left atrium, RA-right atrium, Myo-myocardium, AO-aorta and PA-pulmonary artery). Best viewed in color. (Color figure online)

On the other hand, automatic segmentation and classification of medical images in CHD is rather challenging. Patients with CHD typically suffer from severe variation in heart structures and connections between different parts of the anatomy. Two examples are shown in Fig. 1: the disappearance of the main trunk of pulmonary artery (PA) in (b) (c) introduces much difficulty in the correct segmentation of PA and AO. In addition, CHD does not necessarily cause local tissue changes, as in lesions. As such, hearts with CHD have similar local statistics as normal hearts but with global structural changes. Automatic algorithms to detect the disorders need to be able to capture such changes, which require excellent usage of the contextual information. CHD classification is further complicated by the fact that a patient's CT image may exhibit more than one type of CHD, and the number of types is more than 20 [3].

Various works exist in segmentation and classification of heart with *normal anatomy, e.g.,* [4–6,8,9,13,17,20,22–25, ?], most of which are based on deep neural networks (DNNs) [15, ?]. Recently, researchers started to explore heart segmentation in CHD. The works [7,14,18,19,21] adopt DNNs for blood pool and myocardium segmentation only. The only automatic whole heart and great artery segmentation method in CHD [16] in the literature uses a deep learning and shape similarity analysis based method. A 3D CT dataset for CHD segmentation is also released there. In addition to segmentation, there are also some works about classification of adult heart diseases [2] but not CHD. The automatic classification of CHD still remains a missing piece in the literature due to the complexity of CHD and the lack of dataset.

In this paper, we present ImageCHD, the first medical image dataset for CHD classification. ImageCHD contains 110 3D Computed Tomography (CT) images

which covers 16 types of CHD. CT images are labelled by a team of four experienced cardiovascular radiologists with 7-substructure segmentation and CHD type classification. The dataset is of decent size compared with other medical imaging datasets [17,21]. We also present a baseline method for automatic CHD classification based on the state-of-the-art CHD segmentation framework [16], which is the first automatic CHD classification method in the literature. Results show that the baseline framework can achieve a classification accuracy of 82.0% under selective prediction scheme with 88.4% coverage, and there is still big room for further improvement.
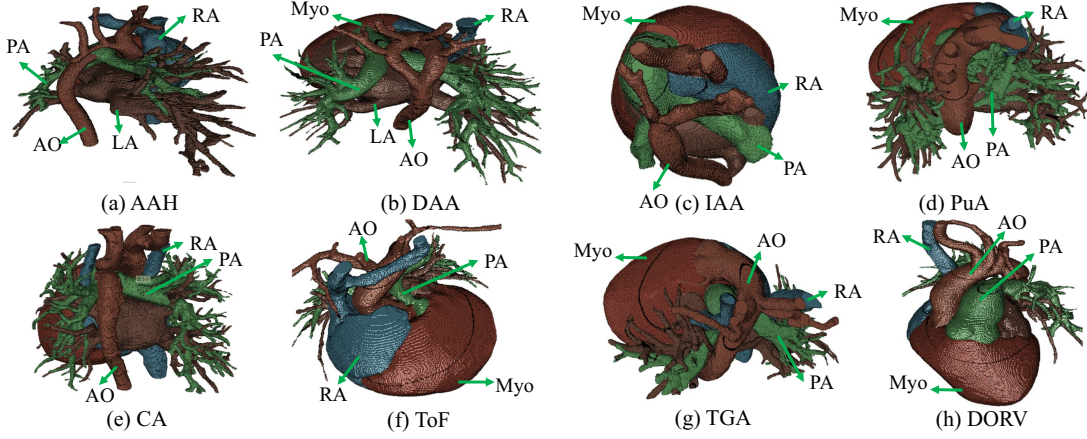

Fig. 2. Examples of CT images in the ImageCHD dataset with its types of CHD.

2 The ImageCHD Dataset

The ImageCHD dataset consists of 3D CT images captured by a Siemens biograph 64 machine from 110 patients, with age 1 month and 40 years (mostly 1 month and 2 years). The size of the images is $512 \times 512 \times (129$–$357)$, and the typical voxel size is $0.25 \times 0.25 \times 0.5\,\mathrm{mm}^3$. The dataset covers 16 types of CHD, which include eight common types (atrial septal defect (ASD), atrioventricular septal defect (AVSD), patent ductus arteriosus (PDA), pulmonary atresia (PuA), ventricular septal defect (VSD), co-arctation (CA), tetrology of fallot (TOF), and transposition of great arteries (TGA)) plus eight less common ones (pulmonary artery sling (PAS), double outlet right ventricle (DORV), common arterial trunk (CAT), double aortic arch (DAA), anomalous pulmonary venous drainage (APVC), aortic arch hypoplasia (AAH), interrupted aortic arch (IAA), double superior vena cava (DSVC)). The number of images associated with each is summarized in Table 1. Several examples of images in the dataset are shown in Fig. 2. Due to the structure complexities, the labeling including segmentation and classification is performed by a team of four cardiovascular radiologists who have extensive experience with CHD. The segmentation label of each image is fulfilled by only one radiologist, and its diagnosis is performed

by four. The time to label each image is around 1–1.5 h on average. The segmentation include seven substructures: LV, RV, LA, RA, Myo, AO and PA.

Table 1. The types of CHD in the ImageCHD dataset (containing 110 3D CT images) and the associated number of images. Note that some images may correspond to more than one type of CHD.

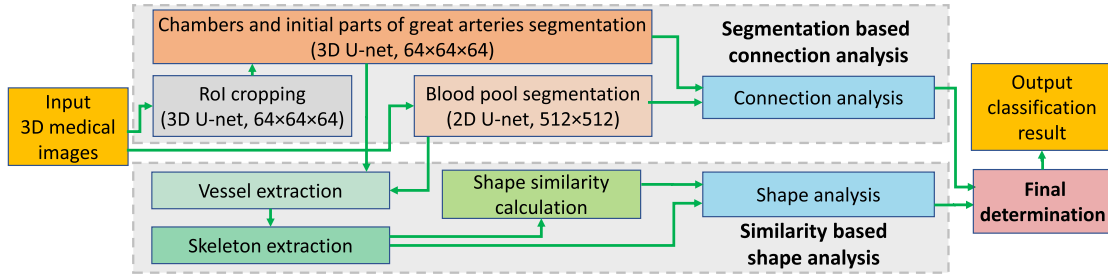| Common CHD								
ASD	AVSD	VSD	TOF	PDA	TGA	CA	PuA	
26	18	44	12	14	7	6	16	
Less Common CHD								Normal
PAS	DORV	CAT	DAA	APVC	AAH	IAA	DSVC	
3	8	4	5	6	3	3	8	6

3   The Baseline Method

Fig. 3. Overview of the baseline method for CHD classification.

Overview: Due to the lack of baseline method for CHD classification, along with the dataset we establish one as shown in Fig. 3, which modifies and extends the whole heart and great artery segmentation method in CHD [16]. It includes two subtasks: segmentation based connection analysis and similarity based shape analysis. Accordingly, the parts and connections most critical to the classification are extracted.

Segmentation Based Connection Analysis: Segmentation is performed with multiple U-Nets [11]. There are two steps in segmentation: blood pool segmentation, and chambers and initial parts of great arteries segmentation. The former is fulfilled by a high-resolution (input size 512×512) 2D U-net, while the latter is performed with a 3D low-resolution (input size $64 \times 64 \times 64$) 3D U-net. A Region of Interest (RoI) cropping is also included with a 3D U-net before the 3D segmentation. With the segmentation results, connection analysis can be
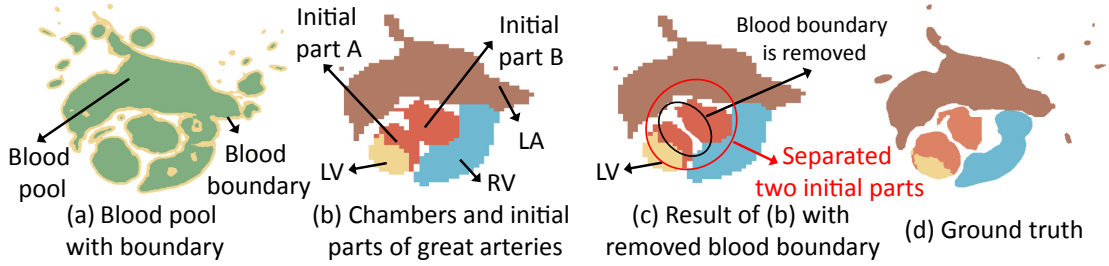
Fig. 4. Connection analysis between LV/RV and great arteries (AO and PA).

processed, which mainly extracts the connection features between great arteries (AO and PA) and LV/RV, and between LV/LA and RV/RA. With the segmentation results, two connection analyses between chambers, AO and PA are then performed by the connection analysis module. The first one analyzes the connections between LV/RV and great arteries. We remove high resolution boundary from low resolution substructures as shown in Fig. 4(a)–(c). Compared with the ground truth in Fig. 4(d), Fig. 4(c) shows that the two initial parts are correctly separated (but not in (b) where they will be treated as connected). The second one has a similar process as the first one.

Similarity Based Shape Analysis: The flow of this subtask is shown in Fig. 5. With the segmentation results, vessel extraction removes the blood pool corresponding to chambers, and vessel refinement removes any remaining small islands in the image, and smooths it with erosion. Then, the skeleton of the vessels are extracted, sampled, normalized, and fed to the shape similarity calculation module to obtain its similarity with all the templates in a pre-defined library. Similarity module is performed using earth mover's distance (EMD) which is a widely used similarity metric for distributions [12]. Two factors need to be modeled: the *weight* of each bin in the distribution, and the *distance* between bins. We model each sampled point in the sampled skeleton as a bin, the Euclidean distance between the points as the distance between bins, and the volume of blood pool around the sampled point as the weight of its corresponding bin. Particularly, the weight is defined as r^3 where r is the radius of the inscribed sphere in the blood pool centered at the sampled point. The template library is manually created in advance and contains six categories of templates corresponding to five types of CHDs and the normal anatomy as shown in Fig. 5, covering all the possible shapes of great arteries in our dataset. Each category contains multiple templates. Finally, the shape analysis module takes the skeleton and its similarities to obtain two kinds of features. The type of the template with the highest similarity is extracted as the first kind. The second kind includes two skeleton features: whether a circle exists in the skeleton, and how the r of the sample points varies. These two features are desired because if there is a circle in the skeleton, the test image is with high possibility to be classified as DAA; If a sampled point with a small r is connected to two sampled points with a much larger r, narrow vessel happens, which is a possible indication of CA and PuA.
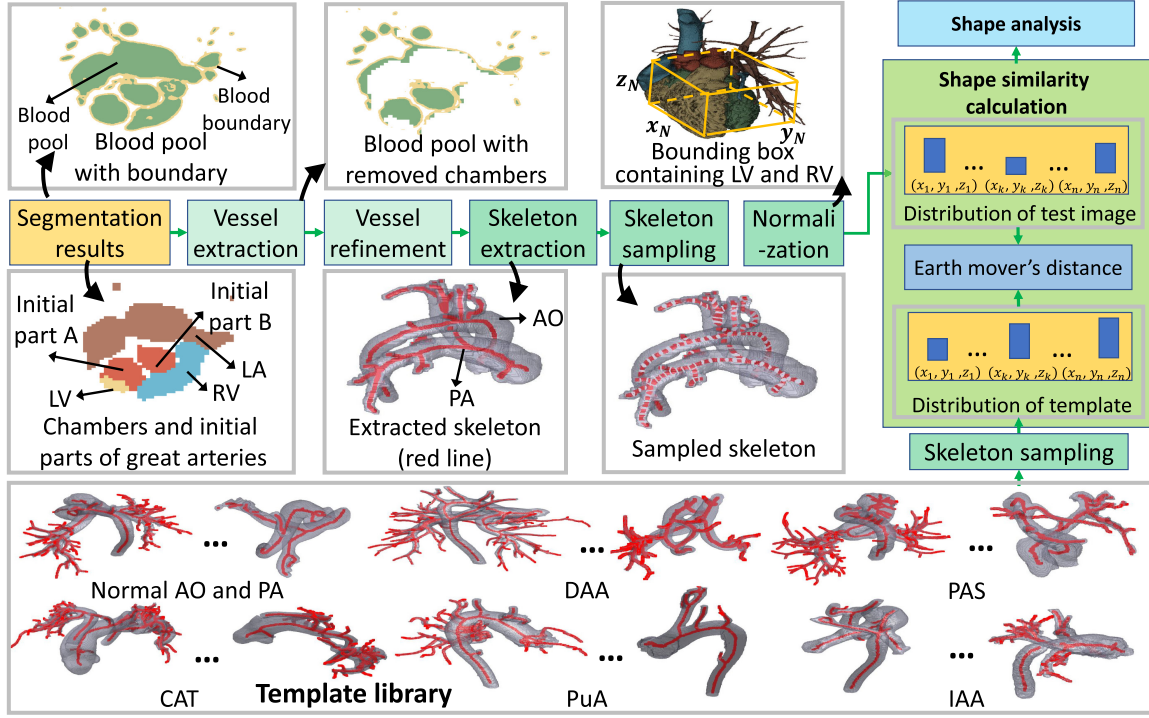
Fig. 5. Similarity based shape analysis of great arteries. Best viewed in color. (Color figure online)

Final Determination: With the extracted connection and shape features, the classification can be finally determined using a rule-based automatic approach. Specifically, ASD and VSD have unexpected connection between LA and RA, and LV and RV, respectively. AVSD is a combination of ASD and VSD, and the three can be classified according to the connection features between LA/LV and RA/RV. DORV has two initial parts of great arteries, both of which are connected to RV. TOF has connected LV and RV, as well as connected LV, RV and the initial part of AO. CHD with specific shapes including CAT, DAA, PuA, PAS and IAA as shown in Fig. 5 can be classified by their shape features. PDA and CA are determined by analyzing the shapes and skeletons such as the variety of r along the skeleton. DSVC can be easily classified by analyzing the skeleton of RV, and APVC is determined by the number of islands that the LA has. Note that if the connection and shape features do not fit any of the above rules, the classifier outputs *uncertain* indicating that the test image cannot be handled and manual classification is needed.

4   Experiment

Experiment Setup: All the experiments run on a Nvidia GTX 1080Ti GPU with 11 GB memory. We implement the 3D U-net using PyTorch based on [8]. For 2D U-net, most configurations remain the same with those of the 3D U-net except that it adopts 5 levels and the number of filters in the initial level is 16. Both Dice loss and cross entropy loss are used, and the training epochs are 2 and 480

for 2D U-net and 3D U-net, respectively. Data augmentation and normalization are also adopted with the same configuration as in [8] for 3D U-net. For both networks and all the analyses, three-fold cross validation is performed (about 37 images for testing, and 73 images for training). We split the dataset such that all types of CHD are present in each subset. The classification considers a total of 17 classes, including 16 types of CHD and the normal anatomy. The templates in the template library are randomly selected from the annotated training set.

In the evaluation, we use selective prediction scheme [10] and report a case as uncertain if at least one chamber is missing (which does not correspond to any type in our dataset) in the segmentation results, or in the similarity calculation the minimum EMD is larger than 0.01. For these cases, manual classification by radiologists is needed. To further evaluate how the baseline method performs against human experts, we also extract manual CT classification from the electronic health records (the manual results can still be wrong).

Results and Analysis: The CHD classification result is shown in Table 2. Each entry (X, Y) in the table corresponds to the number of cases with ground truth class suggested by its row header and predicted class by its column header, where X, and Y are the results from the baseline, and those from radiologists respectively. Again, an image can contribute to multiple cases if it contains more than one types of CHD. From the table we can see that for the baseline method, due to segmentation error or feature extraction failure, 22 cases are classified as uncertain, yielding a 88.4% coverage; Out of the remaining 167 cases, 137 are correct. Thus, for the baseline the overall classification accuracy is 72.5% for full prediction, and 82.0% for selective prediction. For the modified baseline, the overall classification accuracy is 39.2% for full prediction and 50.3% for selective prediction. On the other hand, the manual classification from experienced radiologist can achieve an overall accuracy of 90.5%. It is interesting to note that out of the 17 classes, the baseline method achieves higher accuracy in one (PuA) and breaks even in four (VSD, CAT, DAA, and AAH) compared with manual classification. In addition, Out of the 110 cases, the five radiologists only unanimously agreed on 78 cases, which further reflects the difficulty of the problem and the value of an automated tool.

The mean and standard deviation of Dice score of our baseline method for six substructures of chambers and initial parts of great vessels segmentation, and blood pool segmentation are shown in Table 3. We can notice that blood pool has the highest score, and initial parts of great vessels has the lowest, and the overall segmentation performance is moderate. Though the segmentation performance of initial parts is low, its related types of CHDs (e.g., ToF, TGA) still achieve high classification accuracy which is due to the fact that only the critical segmentation determines the types of CHDs. Comparing the performance of segmentation and classification, we can also notice that accurate segmentation usually helps classification, but not necessarily.

Classification Success: Six types of CHD including TGA, CAT, DAA, AAH, PAS and PuA achieve relatively high accuracy, which is due to their clear and stable features that distinguish them from normal anatomy. Such features can be

Table 2. Number of cases (X, Y) with ground truth class and predicted class suggested by the row and column headers respectively, where X, and Y correspond to automatic classification by the baseline, and manual classification, respectively. Green numbers along the diagonal suggest correct cases. (U-Uncertain, 1-ASD, 2-AVSD, 3-VSD, 4-TOF, 5-PDA, 6-TGA, 7-CA, 8-IAA, 9-PAS, 10-DORV, 11-CAT, 12-DAA, 13-APVC, 14-AAH, 15-PuA, 16-DSVC, N-Normal)

Type	U	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	N
1	6,0	18,24																2,2
2	3,0	1,3	9,14	5,1														
3	1,0	1,2		42,42														
4	1,0		4,2		7,10													
5						7,14												7,0
6	1,0						6,7											
7	1,0							4,6									1,0	
8									2,3									1,0
9	1,0									2,3								
10	1,0		3,1	1,1							3,6							
11												4,4						
12													5,5					
13	1,0													3,6				2,0
14	1,0														2,2			0,1
15	2,0				0,2		0,1						0,1			14,10		
16	1,0																5,7	2,1
N	2,0																	4,6

(a) Poor 3D segmentation due to low data quality
(b) Ground truth of (a)
(c) Error: TOF -> Normal
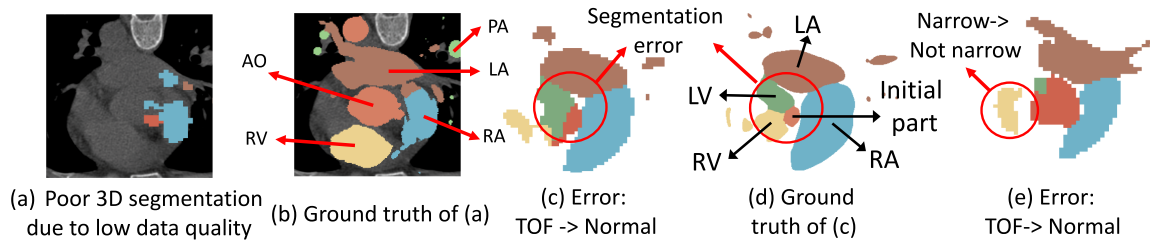(d) Ground truth of (c)
(e) Error: TOF-> Normal

Fig. 6. Examples of classification failure: uncertain classification in (a–b), and wrong classification of TOF in (c) and (e). Best viewed in color. (Color figure online)

easily captured by either connection or shape features extracted by the baseline method. For example, CAT has a main trunk that AO and PA are both connected to; DAA has a circular vessel which is composed of two aortic arches; PAS has a PA with very different shape; PuA has a very thin PA without main trunk; AAH has a long period of narrow vessels in the arch; and TGA has a reversed connection to LV and RV.

Classification Failure: Test images are classified as uncertain due to segmentation error. Figure 6 shows some examples of such error. The test image in Fig. 6(a) has very low contrast, and its blood pool and boundary are not clear compared with other areas, resulting in segmentation error: compared with the

Table 3. Mean and standard deviation of Dice score of our baseline method (in %) for six substructures of chambers and initial parts of great vessels segmentation, and blood pool segmentation.

LV	RV	LA	RA	Initial parts of great vessels	Blood pool	Average
77.7 ± 16.2	74.6 ± 13.8	77.9 ± 11.2	81.5 ± 11.5	66.5 ± 15.1	86.5 ± 10.5	75.6 ± 10.2

ground truth in Fig. 6(b), only RA and part of the initial parts of great arteries are segmented. As for the cases where a CHD type is predicted but wrong, we will use TOF as examples, and leave the comprehensive discussion for all classes in the supplementary material. Segmentation error around the initial parts of great arteries is the main reason of the classification failure of TOF as shown in Fig. 6. Compared with the ground truth in Fig. 6(d), the 3D segmentation in Fig. 6(c) labels part of LV as RV, resulting in the initial part only connected to RV rather than RV and LV. As one of the main features of TOF is that one initial part is connected to both RV and LV, missing such feature leads to misclassification of TOF as VSD. Another main feature of TOF is the narrow vessels in the initial part and its connected RV part, which can also lead to wrong classification if not detected correctly as shown in Fig. 6(e). A precise threshold to decide whether the vessels are narrow or not is still missing even in clinical studies.

Discussion: We can notice that segmentation accuracy is important for successful classification of CHD. Higher segmentation accuracy can lead to better connection and shape feature extraction. In addition, so far we have only considered the connection features in the blood pool and the shapes of the vessels. More structural features associated with classification should be considered to improve the performance, which due to the lack of local tissue changes, need innovations from the deep learning community and deeper collaboration between computer scientists and radiologists.

5 Conclusion

We introduce to the community the ImageCHD dataset [1] in hopes of encouraging new research into unique, difficult and meaningful datasets. We also present a baseline method for comparison on this new dataset, based on a state-of-the-art whole-heart and great artery segmentation method for CHD images. Experimental results show that under selective prediction scheme the baseline method can achieve a classification accuracy of 82.0%, leaving big room for improvement. We hope that the dataset and the baseline method can encourage new research that be used to better address not only the CHD classification but also a wider class of problems that have large global structural change but little local texture/feature change.

References

1. https://github.com/XiaoweiXu/ImageCHD-A-3D-Computed-Tomography-Image-Dataset-for-Classification-of-Congenital-Heart-Disease
2. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
3. Bhat, V., BeLaVaL, V., Gadabanahalli, K., Raj, V., Shah, S.: Illustrated imaging essay on congenital heart diseases: multimodality approach Part I: clinical perspective, anatomy and imaging techniques. J. Clin. Diagn. Res. JCDR **10**(5), TE01 (2016)
4. Dou, Q., et al.: PnP-AdaNet: plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access **7**, 99065–99076 (2019)
5. Habijan, M., Leventić, H., Galić, I., Babin, D.: Whole heart segmentation from CT images using 3D U-net architecture. In: 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 121–126. IEEE (2019)
6. Liu, T., Tian, Y., Zhao, S., Huang, X., Wang, Q.: Automatic whole heart segmentation using a two-stage U-net framework and an adaptive threshold window. IEEE Access **7**, 83628–83636 (2019)
7. Pace, D.F., et al.: Iterative segmentation from limited training data: applications to congenital heart disease. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 334–342. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_38
8. Payer, C., Štern, D., Bischof, H., Urschler, M.: Multi-label whole heart segmentation using CNNs and anatomical label configurations. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 190–198. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_20
9. Piccini, D., Littmann, A., Nielles-Vallespin, S., Zenge, M.O.: Respiratory self-navigation for whole-heart bright-blood coronary MRI: methods for robust isolation and automatic segmentation of the blood pool. Magn. Reson. Med. **68**(2), 571–579 (2012)
10. Pidan, D., El-Yaniv, R.: Selective prediction of financial trends with hidden Markov models. In: Advances in Neural Information Processing Systems, pp. 855–863 (2011)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vision **40**(2), 99–121 (2000)
13. Wang, C., MacGillivray, T., Macnaught, G., Yang, G., Newby, D.: A two-stage 3D Unet framework for multi-class segmentation on full resolution image. arXiv preprint arXiv:1804.04341 (2018)
14. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease. In: Zuluaga, M.A., Bhatia, K., Kainz, B., Moghari, M.H., Pace, D.F. (eds.) RAMBO/HVSMR-2016. LNCS, vol. 10129, pp. 95–102. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52280-7_9

15. Xu, X., et al.: Quantization of fully convolutional networks for accurate biomedical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8300–8308 (2018)
16. Xu, X., et al.: Whole heart and great vessel segmentation in congenital heart disease using deep neural networks and graph matching. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 477–485. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_53
17. Xu, Z., Wu, Z., Feng, J.: CFUN: combining faster R-CNN and U-net network for efficient whole heart segmentation. arXiv preprint arXiv:1812.04914 (2018)
18. Yang, X., Bian, C., Yu, L., Ni, D., Heng, P.-A.: Class-balanced deep neural network for automatic ventricular structure segmentation. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 152–160. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_16
19. Yang, X., Bian, C., Yu, L., Ni, D., Heng, P.-A.: Hybrid loss guided convolutional networks for whole heart parsing. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 215–223. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_23
20. Ye, C., Wang, W., Zhang, S., Wang, K.: Multi-depth fusion network for whole-heart CT image segmentation. IEEE Access **7**, 23421–23429 (2019)
21. Yu, L., Yang, X., Qin, J., Heng, P.-A.: 3D FractalNet: dense volumetric segmentation for cardiovascular MRI volumes. In: Zuluaga, M.A., Bhatia, K., Kainz, B., Moghari, M.H., Pace, D.F. (eds.) RAMBO/HVSMR-2016. LNCS, vol. 10129, pp. 103–110. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52280-7_10
22. Zhang, R., Chung, A.C.S.: A fine-grain error map prediction and segmentation quality assessment framework for whole-heart segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 550–558. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_61
23. Zheng, H., et al.: HFA-Net: 3D cardiovascular image segmentation with asymmetrical pooling and content-aware fusion. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 759–767. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_84
24. Zhou, Z., et al.: Cross-modal attention-guided convolutional network for multi-modal cardiac segmentation. In: Suk, H.-I., Liu, M., Yan, P., Lian, C. (eds.) MLMI 2019. LNCS, vol. 11861, pp. 601–610. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32692-0_69
25. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. **31**, 77–87 (2016)