Title

Automating Cardiac Segmentation for Complex Congenital Heart Disease 3D Modeling Using nnU-Net

ISBN

9798189294495

Author

Outtrim, Connor Erickson

Publication Date

2026-09-03

Peer reviewed|Thesis/dissertation

Automating Cardiac Segmentation for Complex Congenital Heart Disease 3D Modeling Using nnU-Net

by

Connor Outtrim

THESIS

Submitted in partial satisfaction of the requirements for degree of

MASTER OF SCIENCE

in

Biomedical Imaging

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

Shafkat Anwar

69B4E00D907344D...

Shafkat Anwar

Chair

DocuSigned by:

Alastair Martin

Alastair Martin

DocuSigned by:

David Saloner

David Saloner

DocuSigned by:

Youngho Seo

C525AA918D85484...

Youngho Seo

Committee Members

Acknowledgments

I would like to first express my gratitude to my committee chair, Dr. Shafkat Anwar. Your optimistic outlook, intellectual curiosity, and enthusiasm for innovation and problem-solving set an admirable example that I hope to carry forward throughout my career in medicine. Dr. Sagar Jalui, thank you for your insightful guidance throughout this project. Your mentorship pushed me to grow beyond the technical hurdles of this work and gave me the confidence to work through challenges that once seemed beyond my abilities. To Michael Bunker, thank you for your support throughout this project and for your expertly annotated segmentations, which have been the backbone of my work. To my committee members, Dr. David Saloner, Dr. Alastair Martin, and Dr. Youngho Seo, you are outstanding professors and remarkable scientists. I sincerely thank you for your fascinating classes and for sharing your expertise in biomedical imaging.

To the MSBI program, thank you for creating such a unique environment that encourages excellence in both educational and professional endeavors. The faculty, staff, and students truly form an exceptional community, and I am grateful to have learned alongside such brilliant minds.

I would like to give a special thank you to my family. Mom and Dad, I am grateful for your unwavering support, which has allowed me to lead a life full of curiosity, passion, and love. Thank you to my brother, Ryan, and my sister, Sophie, who have always been part of the reason I strive to become someone they can be proud of. To my partner, Amanda Luu, thank you for your patience and encouragement throughout this year. Because of you, I never forgot to eat well and take care of myself.

Automating Cardiac Segmentation for Complex Congenital Heart Disease 3D Modeling Using nnU-Net

Connor Outtrim

Abstract

Patient-specific three-dimensional (3D) models can support surgical planning for complex congenital heart disease (CHD), but manual segmentation of cardiac computed tomography (CT) is time intensive. This thesis evaluated nnU-Net v2 for multi-structure CHD CT segmentation in three phases. Phase 1 screened nnU-Net configurations on the public CHD68 dataset and selected a 1,250-epoch model with learning rate 0.01. The resulting CHD68 model (M0) achieved mean case-level Dice 0.846 on 14 validation cases. Phase 2 applied M0 without additional training to 41 UCSF CT cases to establish a local baseline. Across six shared blood-pool structures, mean Dice decreased from 0.845 on CHD68 to 0.389 on UCSF, with substantial case heterogeneity and several complete structure omissions. Phase 3 trained a UCSF blood-pool model (M1), a UCSF + CHD68 blood-pool model (M2), and a ten-label UCSF hollow model (M3). On the same nine held-out UCSF blood-pool cases, mean Dice was 0.488 for M0, 0.754 for M1, and 0.755 for M2; M2 also largely retained CHD68 performance. M3 achieved mean Dice 0.469 across ten structures and 14 validation cases. A secondary ground-truth-bounded analysis increased mean Dice to 0.484 (+0.015), with the largest changes in the SVC, aorta, IVC, and pulmonary veins. Qualitative review showed that several nonfailure hollow predictions preserved major cardiac geometry despite modest overlap. Hollow and blood-pool Dice are not directly comparable because the targets differ in thickness, topology, branch extent, and truncation. The findings support continued development of hollow segmentation as an expert-editable starting point for the CA3D+ 3D-modeling workflow, with annotation refinement, cohort expansion, and direct evaluation of correction burden and printability as next steps.

Table of Contents

Chapter 1 Introduction 1

1.1 Clinical context of congenital heart disease 1

1.2 CT-based 3D modeling and surgical planning 1

1.3 Manual segmentation as a workflow bottleneck 2

1.4 Deep learning segmentation and nnU-Net 2

1.5 Prior CHD segmentation work and external generalization 3

1.6 Study objectives and hypotheses 4

Chapter 2 Materials and Methods 6

2.1 Overall experimental design 6

2.2 Ethics, data governance, and de-identification 7

2.3 Computing environment and software 7

2.4 Phase 1 public dataset and preprocessing 8

2.5 Phase 1 configuration screening and model selection 9

2.6 UCSF cases and annotation provenance 9

2.7 Phase 2 cohort and quality control 10

2.8 Phase 2 preprocessing, inference, and label mapping 10

2.9 Phase 3 training cohorts and model development 11

2.10 Evaluation metrics 12

2.11 Statistical analysis 12

2.12 Ground-truth-bounded hollow analysis 13

2.13 Qualitative evaluation 13

2.14 Exploratory Phase 2 case-characteristic analysis 14

Chapter 3 Results 15

3.1 Phase 1 configuration selection and validation performance 15
3.2 Phase 2 performance on UCSF data 16
3.3 Structure-level Phase 1 versus Phase 2 performance 17
3.4 Phase 2 failure patterns and case heterogeneity 18
3.5 Phase 3 blood-pool comparison on nine UCSF validation cases 19
3.6 Phase 3 blood-pool performance by structure 20
3.7 M2 performance on CHD68 validation cases 21
3.8 Phase 3 ten-label hollow segmentation 21
3.9 Ground-truth-bounded hollow analysis 23
3.10 Qualitative evaluation of Phase 3 validation cases 24
Chapter 4 Discussion 27
4.1 Principal findings and hypothesis resolution 27
4.2 Phase 1 configuration screening 28
4.3 Performance on UCSF data and cross-dataset generalization 28
4.4 UCSF-trained blood-pool models 29
4.5 Hollow segmentation and 3D-modeling relevance 29
4.6 Annotation consistency and workflow implications 30
4.7 Limitations 31
4.8 Future directions 32
References 34

List of Figures

Figure 2.1. Study overview. 7

Figure 3.1. Phase 1 learning-rate screening. 15

Figure 3.2. Phase 1 versus Phase 2 shared-structure Dice distributions. 17

Figure 3.3. Phase 3 blood-pool comparison on nine UCSF validation cases. 20

Figure 3.4. M3 hollow segmentation performance by structure. 23

Figure 3.5. Change in M3 Dice after ground-truth bounding. 24

Figure 3.6. Representative 2D Phase 3 segmentation examples. 25

Figure 3.7. Representative 3D blood-pool and hollow segmentations. 26

List of Tables

Table 1.1. Selected prior congenital-heart segmentation studies relevant to this thesis. 4

Table 2.1. Final three-phase study design and model nomenclature. 6

Table 2.2. Software and computing environment. 8

Table 2.3. CHD68 labels and Phase 2 six-label evaluation mapping. 11

Table 2.4. Phase 3 training and validation datasets. 12

Table 3.1. Selected Phase 1 performance at learning rate 0.01 and 1,250 epochs (n = 14 CHD68 validation cases). 16

Table 3.2. Structure-level Phase 1 versus Phase 2 Dice comparisons for the six shared blood-pool structures. 17

Table 3.3. Phase 2 segmentation metrics and prediction failures (n = 41 UCSF cases). 18

Table 3.4. Paired case-level comparisons on nine UCSF blood-pool validation cases. 19

Table 3.5. Mean Dice by structure on the same nine UCSF validation cases. 20

Table 3.6. CHD68 six-structure means for M0 and M2 on the same 14 validation cases. 21

Table 3.7. M3 ten-label hollow performance by structure (n = 14 UCSF validation cases). 22

List of Abbreviations

Abbreviation	Definition
Ao	Aorta
CA3D+	Center for Advanced 3D+ Technologies
CHD	Congenital heart disease
CT	Computed tomography
DICOM	Digital Imaging and Communications in Medicine
FOV	Field of view
GT	Ground truth
HD95	95th-percentile Hausdorff distance
HU	Hounsfield units
IoU	Intersection over union
IVC	Inferior vena cava
LA	Left atrium
LV	Left ventricle
MRI	Magnetic resonance imaging
Myo	Myocardium
PA	Pulmonary artery
PV	Pulmonary veins

Abbreviation	Definition
RA	Right atrium
RAS	Right-anterior-superior orientation
RV	Right ventricle
SVC	Superior vena cava
UCSF	University of California, San Francisco

Chapter 1
Introduction

1.1 Clinical context of congenital heart disease

Congenital heart disease (CHD) affects approximately 0.9% of live births worldwide [1]. The term encompasses a broad spectrum of structural abnormalities, from isolated septal defects to complex combinations of abnormal chamber morphology, great-vessel connections, venous return, and single-ventricle physiology. Many patients with complex CHD undergo staged operations early in life, and the anatomy encountered during surgical planning may differ substantially from normal cardiac organization.

The anatomical diversity of CHD creates a visualization problem. Standard axial, coronal, and sagittal images contain the required information, but unusual spatial relationships can be difficult to integrate mentally, particularly when chambers, vessels, conduits, and postoperative pathways intersect. This problem motivates patient-specific three-dimensional representation.

1.2 CT-based 3D modeling and surgical planning

Cardiac computed tomography (CT) provides high-resolution volumetric information that can be reformatted in multiple planes and converted into patient-specific three-dimensional (3D) models. These models can support communication, procedural planning, and surgical preparation by making abnormal chamber and vessel relationships easier to inspect. At the UCSF Center for Advanced 3D+ Technologies (CA3D+), CT-derived models are incorporated into a workflow that includes image review, anatomical segmentation, surface construction, quality control, and 3D visualization or printing.

A blood-pool segmentation is not identical to the hollow representation used for the final CA3D+ model. Blood-pool masks define the contrast-filled lumen of cardiac chambers and vessels. Hollow segmentations instead represent wall- and surface-like anatomy. CA3D+ uses

hollow representations because they more closely approximate the surfaces a surgeon encounters and provide shell-like geometry that is better suited to physical 3D printing. Creating a hollow target also requires explicit decisions about wall thickness, branch inclusion, vessel extent, junction boundaries, and truncation planes. These properties make hollow overlap more sensitive to small boundary or termination differences than filled blood-pool overlap, so the two Dice values should be interpreted within their respective target definitions.

1.3 Manual segmentation as a workflow bottleneck

The clinical value of a 3D model depends on a reliable segmentation of the relevant anatomy. Within the CA3D+ workflow, cardiac chambers and vessels are segmented manually or semi-automatically in commercial software, and complex anatomy, weak contrast boundaries, small vessels, prior surgical alterations, and overlapping structures can require substantial expert correction. This segmentation and correction burden limits the speed and scalability of on-demand 3D model creation.

An automated model does not need to replace expert review to provide value. A useful system could generate an initial segmentation that is then checked and corrected by a trained operator. The aim of this thesis was therefore to develop and evaluate nnU-Net for CHD CT segmentation, establish how a model developed on public data performed on UCSF cases, and investigate blood-pool and hollow models trained with UCSF data.

1.4 Deep learning segmentation and nnU-Net

Deep convolutional networks have achieved strong performance in many medical-image segmentation tasks. Reported results, however, can depend heavily on preprocessing, patch size, normalization, augmentation, network depth, and training schedule. nnU-Net was developed to reduce this configuration burden by deriving a training plan from the dataset fingerprint rather

than requiring a manually designed architecture [2]. Controlled benchmarking has continued to support nnU-Net as a strong reference method for three-dimensional medical segmentation [3].

The self-configuring design of nnU-Net was well suited to this project because it provided a consistent framework for configuration screening, evaluation on a separate clinical dataset, retraining with UCSF cases, and extension from conventional blood-pool targets to the hollow representation used in the CA3D+ workflow.

1.5 Prior CHD segmentation work and external generalization

Automated cardiac CT segmentation has been studied more extensively in normal or adult anatomy than in congenital disease. A 2024 systematic review reported median Dice similarity coefficients of 0.83 to 0.92 for commonly segmented cardiac structures on CT, but 81% of included studies did not test or validate their models on external data and ground-truth reporting was inconsistent [4]. In CHD, Xu et al. developed a deep-learning and graph-matching framework using 68 three-dimensional CT studies [5], and Yao et al. later reported Dice coefficients of 0.826 for simple CHD and 0.741 for complex CHD [6]. Zhu et al. reported a whole-heart Dice coefficient of 0.869 on the ImageCHD dataset using a hybrid encoder-decoder model in 2025 [7]. nnU-Net v2 has also been used for multi-label fetal cardiac ultrasound segmentation, demonstrating that it can learn numerous small structures in a related cardiac application [8].

The literature reviewed for this thesis did not identify a published evaluation of an nnU-Net CHD CT model on an independent clinical cohort. This distinction matters because strong performance on a held-out portion of one dataset does not guarantee similar performance on cases collected and annotated elsewhere. Differences in patient scale, anatomy, scanner acquisition, contrast timing, field of view, and annotation conventions can all affect cross-dataset

generalization. Testing a public-data model on UCSF cases therefore provided a clinically relevant baseline before models were trained with UCSF data.

Table 1.1. Selected prior congenital-heart segmentation studies relevant to this thesis.

Study	Data	Method	Labels	Reported result
Xu et al., 2019	68 CHD CT	Deep network + graph matching	7	Public CHD dataset and whole-heart/great-vessel framework
Yao et al., 2023	68 CHD CT	2D + 3D U-Net + graph matching	7	Dice 0.826 simple; 0.741 complex CHD
Zhu et al., 2025	CHD CT/MRI cohort	Hybrid encoder-decoder	7	Whole-heart Dice 0.869 on ImageCHD
Liang et al., 2024	Fetal cardiac ultrasound	nnU-Net v2	15	Mean Dice 0.871

1.6 Study objectives and hypotheses

This thesis was organized around three linked objectives. Phase 1 screened nnU-Net configurations using the public CHD68 dataset and selected a reproducible model configuration. Phase 2 applied that model to UCSF CHD CT cases without additional training to establish baseline performance on the local clinical dataset and characterize failure patterns. Phase 3 used UCSF data to develop blood-pool models and a ten-label hollow model aligned with the CA3D+ 3D-modeling workflow. Blood-pool segmentation served as the more conventional reference task represented in prior cardiac-segmentation literature, whereas the hollow experiment addressed the representation used locally for downstream modeling and printing.

The Phase 1 hypothesis was that nnU-Net could achieve mean Dice greater than 0.80 on held-out CHD68 cases without architecture customization. The Phase 2 hypothesis was that performance would decrease when M0 was applied to UCSF cases without additional training. The Phase 3 blood-pool hypothesis was that training with UCSF cases would improve performance on held-

out UCSF cases relative to the Phase 1 model. The mixed UCSF + CHD68 experiment additionally tested whether that improvement could coexist with retention of CHD68 performance. The hollow experiment assessed the feasibility of learning the expanded ten-label target used for CA3D+ modeling.

Chapter 2
Materials and Methods

2.1 Overall experimental design

This project used a three-phase design. Phase 1 screened learning rate and training duration on the public CHD68 dataset and selected a reproducible nnU-Net v2 configuration; the selected CHD68 model is referred to as M0. Phase 2 applied M0 to UCSF CT cases without additional training to establish baseline performance on the local clinical dataset. Phase 3 used the selected configuration to train three models from scratch: the UCSF blood-pool model (M1), the UCSF + CHD68 blood-pool model (M2), and the UCSF hollow model (M3). Each model was trained using a predefined training set and evaluated on a held-out validation set.

The phases answered separate questions: whether nnU-Net could perform well on the public CHD68 dataset, how the selected model performed on UCSF cases as a baseline, and what performance could be achieved after training with UCSF blood-pool or hollow segmentations. M2 additionally tested whether a model trained with both datasets could retain CHD68 performance while learning from UCSF cases.

Table 2.1. Final three-phase study design and model nomenclature.

Phase / model	Training data	Held-out evaluation data	Primary purpose
Phase 1 / M0: CHD68 model	CHD68: 54 cases	CHD68: 14 cases	Configuration screening and public-dataset reference performance
Phase 2 / M0 on UCSF	No additional training	UCSF: 41 CT cases	Establish baseline performance on UCSF clinical data
Phase 3 / M1: UCSF blood-pool	UCSF: 32 cases	UCSF: 9 cases	Develop a six-label UCSF blood-pool model
Phase 3 / M2: UCSF + CHD68 blood-pool	UCSF + CHD68: 86 cases	9 UCSF + 14 CHD68 cases	Develop a mixed-data blood-pool model and assess CHD68 retention

Phase / model	Training data	Held-out evaluation data	Primary purpose
Phase 3 / M3: UCSF hollow	UCSF hollow: 58 cases	UCSF hollow: 14 cases	Develop the ten-label hollow target used for CA3D+ modeling

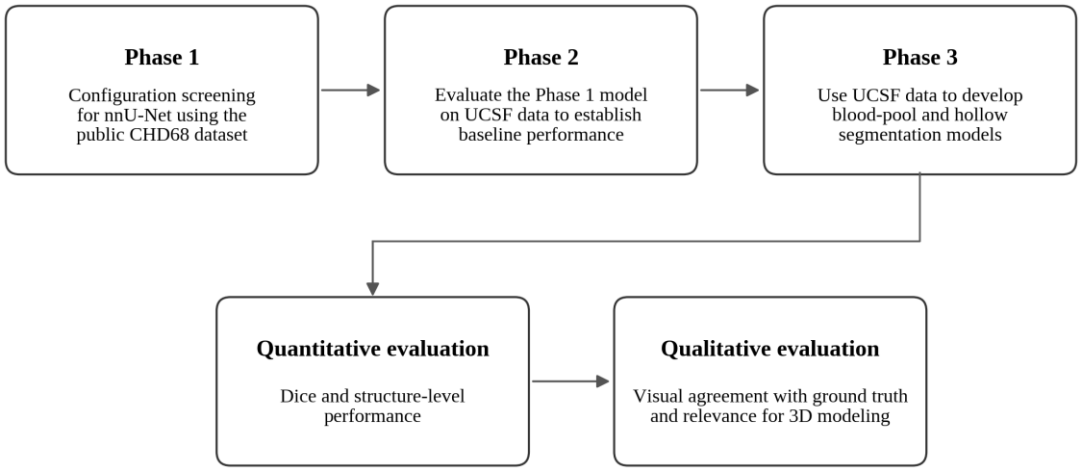


Figure 2.1. Study overview.
Phase 1 screened nnU-Net configurations using the public CHD68 dataset. Phase 2 applied the selected Phase 1 model to UCSF CT cases to establish baseline performance. Phase 3 used UCSF data to develop blood-pool and hollow segmentation models; one blood-pool experiment also incorporated CHD68 cases. Final performance was assessed quantitatively and qualitatively, including visual agreement relevant to the intended 3D-modeling workflow.

2.2 Ethics, data governance, and de-identification

The UCSF portion of the study used previously collected clinical congenital-heart CT studies and segmentation exports maintained within the CA3D+ workflow. Protected health information was removed before the images and segmentation files were used for this project. Analytic outputs and thesis figures contain de-identified case identifiers only.

2.3 Computing environment and software

Model training and inference were performed on a Windows 11 workstation with an NVIDIA GeForce RTX 5080 Laptop GPU (16 GB). The modeling environment used nnU-Net v2 with

PyTorch and CUDA. Software and hardware versions used for training, data preparation, analysis, clinical segmentation review, and qualitative visualization are summarized in Table 2.2.

Table 2.2. Software and computing environment.

Component	Version / configuration	Use
Operating system	Windows 11 25H2; build 10.0.26200.0	Training, inference, and data preparation
GPU / driver	NVIDIA GeForce RTX 5080 Laptop GPU (16 GB); driver 592.02	Model training and inference
Python	3.14.4; dedicated virtual environment	Model and analysis environment
PyTorch / CUDA / cuDNN	PyTorch 2.11.0+cu130; CUDA 13.0; cuDNN 9.19.0	Model training and inference
nnU-Net	v2.6.4	Planning, preprocessing, training, and prediction
Analysis packages	NumPy 2.4.3; SciPy 1.17.1; pandas 3.0.2; nibabel 5.4.2; SimpleITK 2.5.3; pydicom 3.0.2; Matplotlib 3.10.8	Data preparation, metrics, statistics, and figures
Materialise Mimics Core Medical	27.0	Clinical segmentation review and export
Materialise 3-matic Medical	19.0	Surface and hollow-model workflow
3D Slicer	5.10.0	Qualitative review and overlay visualization

2.4 Phase 1 public dataset and preprocessing

Phase 1 used the public CHD68 dataset released with the whole-heart and great-vessel segmentation work of Xu et al. [5]. The dataset contains 68 three-dimensional CT volumes; 54 cases were used for training and 14 were held out for validation. Seven foreground structures were modeled: left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium (Myo), aorta (Ao), and pulmonary artery (PA).

The CHD68 images used a nonnegative stored-value CT representation ranging approximately from 0 to 4095 and were represented in right-anterior-superior (RAS) orientation. These

characteristics were used as compatibility references when preparing UCSF images for Phase 2 inference.

2.5 Phase 1 configuration screening and model selection

Three learning rates were evaluated: 0.1, 0.01, and 0.001. Runs used the nnU-Net v2 3d_fullres configuration with Dice plus cross-entropy loss. Intermediate checkpoints were retained at planned training milestones, and selected runs were extended through 1,500 epochs to confirm that performance had converged.

The final configuration was selected using the 14 held-out CHD68 validation cases before Phase 2 evaluation. The selected configuration used learning rate 0.01 and the 1,250-epoch checkpoint. Extending training to 1,500 epochs did not materially improve validation pseudo-Dice. Training progress and checkpoint selection were monitored using nnU-Net's validation pseudo-Dice, while final model performance was calculated separately using the evaluation metrics described in Section 2.10. The same 3d_fullres configuration, Dice plus cross-entropy loss, learning rate 0.01, and 1,250 training epochs were used for the three Phase 3 training runs.

2.6 UCSF cases and annotation provenance

The UCSF cases used in Phases 2 and 3 consisted of clinical congenital-heart CT studies and segmentation exports maintained by the Center for Advanced 3D+ Technologies (CA3D+). The archive contained both blood-pool and hollow anatomical masks. UCSF reference segmentations were generated by experienced CA3D+ medical 3D engineers under supervision of the principal investigator, a senior cardiologist with expertise in advanced cardiac imaging. Segmentation production and review used Materialise Mimics Core Medical and 3-matic Medical.

For blood-pool comparison, evaluation was restricted to the six structures shared between CHD68 and the UCSF blood-pool exports: LV, RV, LA, RA, Ao, and PA. Myocardium was

omitted because corresponding UCSF ground truth was not consistently available. The Phase 3 hollow target contained ten UCSF structures: Ao, coronary arteries, inferior vena cava (IVC), LA, LV, PA, pulmonary veins (PV), RA, RV, and superior vena cava (SVC). Blood-pool and hollow annotations were treated as distinct segmentation targets.

2.7 Phase 2 cohort and quality control

Phase 2 included 41 UCSF CT cases with valid ground-truth masks for the six shared blood-pool structures: LV, RV, LA, RA, Ao, and PA. Before evaluation, cases underwent quality-control review for CT availability, label completeness, image-mask geometry, and major annotation defects. The CHD68 model (M0) also predicted myocardium, but myocardium was not scored because corresponding UCSF ground truth was not consistently available.

2.8 Phase 2 preprocessing, inference, and label mapping

To match the input representation of the CHD68 dataset used to train M0, UCSF CT volumes and corresponding masks were reoriented to RAS, and CT intensities were converted to the same nonnegative stored-value representation used by CHD68. These preprocessing steps were applied consistently to all 41 Phase 2 cases before inference.

Phase 2 then applied M0, the selected 1,250-epoch CHD68 checkpoint, without additional training or fine-tuning. Inference used the nnU-Net 3d_fullres configuration with the standard inference settings. For six-label UCSF evaluation, LV, RV, LA, and RA retained labels 1 through 4; myocardium was mapped to background; Ao was remapped from CHD68 label 6 to evaluation label 5; and PA was remapped from label 7 to evaluation label 6.

Table 2.3. CHD68 labels and Phase 2 six-label evaluation mapping.

Structure	CHD68 source label	Phase 2 evaluation label	Scored in Phase 2
LV	1	1	Yes
RV	2	2	Yes
LA	3	3	Yes
RA	4	4	Yes
Myo	5	0 (background)	No; no corresponding UCSF ground truth
Ao	6	5	Yes
PA	7	6	Yes

2.9 Phase 3 training cohorts and model development

Phase 3 used the selected Phase 1 configuration to train three models from scratch. The models differed in training data and target definition but otherwise used nnU-Net v2 3d_fullres, Dice plus cross-entropy loss, learning rate 0.01, and 1,250 training epochs.

M1, the UCSF blood-pool model, used 41 UCSF cases with the six shared blood-pool labels: 32 for training and nine for validation. M2, the UCSF + CHD68 blood-pool model, used the 41 UCSF blood-pool cases together with all 68 CHD68 cases after both datasets were mapped to the same six-label space. Of the 109 total cases, 86 were used for training and 23 for validation. The M2 validation set contained the same nine UCSF cases used for M1 and the same 14 CHD68 validation cases used in Phase 1, permitting paired comparisons on each dataset.

M3, the UCSF hollow model, used 72 UCSF cases with the ten-label hollow target. Fifty-eight cases were used for training and 14 were held out for validation. The labels were Ao, coronary arteries, IVC, LA, LV, PA, PV, RA, RV, and SVC. This ten-label experiment was the primary hollow endpoint for the thesis.

Table 2.4. Phase 3 training and validation datasets.

Model	Target space	Training, n	Validation, n	Validation composition
M1: UCSF blood-pool	6 labels	32	9	9 UCSF cases
M2: UCSF + CHD68 blood-pool	6 labels	86	23	9 UCSF + 14 CHD68 cases
M3: UCSF hollow	10 labels	58	14	14 UCSF cases

2.10 Evaluation metrics

Segmentation overlap was quantified primarily with the Dice similarity coefficient and intersection over union (IoU). For predicted mask P and ground-truth mask G, Dice = $2|P \cap G| / (|P| + |G|)$, and IoU = $|P \cap G| / |P \cup G|$. Precision and recall were calculated from true-positive, false-positive, and false-negative voxels. Boundary error was summarized with the 95th-percentile Hausdorff distance (HD95) in millimeters.

HD95 was calculated as the 95th percentile of the symmetric surface-distance distribution after retaining the largest connected component of the predicted structure while leaving the complete ground-truth structure unchanged. HD95 and precision are undefined for empty predictions, so empty predictions were reported separately. Because HD95 distributions were long-tailed, median and interquartile range were used for boundary-error summaries where appropriate.

Phase 2 overlap metrics compared each predicted structure with its original independent binary UCSF ground-truth mask. These masks could overlap anatomically. Phase 3 blood-pool metrics were calculated in the common six-label representation, whereas M3 was evaluated in its separate ten-label hollow target.

2.11 Statistical analysis

Phase 1 and Phase 2 were compared as independent cohorts using the six shared blood-pool structures. Structure-level differences were tested with two-sided Mann-Whitney U tests, with

Holm correction across the six structures. Cliff's delta quantified effect size, with negative values indicating lower Phase 2 performance. Ten thousand case-level nonparametric bootstrap samples were used to estimate 95% confidence intervals for overall mean and median differences.

Phase 3 blood-pool comparisons used paired inference because M0, M1, and M2 were evaluated on the same nine UCSF validation cases. Paired differences in case-level mean Dice were summarized with nonparametric bootstrap 95% confidence intervals and exact sign-flip permutation tests. Permutation-test p values were adjusted using the Holm method across the three pairwise model comparisons. Structure-level results were summarized descriptively. M2 was also evaluated on the same 14 CHD68 validation cases used in Phase 1 to assess retention of CHD68 performance.

2.12 Ground-truth-bounded hollow analysis

A secondary ground-truth-bounded analysis was performed for M3 to estimate how much Dice was affected by prediction extent beyond the annotated anatomy. For each of the 14 hollow validation cases, a single axis-aligned three-dimensional bounding box was defined from the union of all ten ground-truth labels with no additional padding. Prediction voxels outside this box were set to background, while labels inside the box were unchanged, and Dice was recomputed. Because ground truth defined the allowable extent, this was an optimistic analysis intended only to estimate the contribution of truncation and distal prediction extent; it was not a deployable postprocessing method.

2.13 Qualitative evaluation

Qualitative review was used to illustrate the range of Phase 3 model performance. Worst, median, and best cases by case-level mean Dice were selected for each model. Axial images used

solid green ground-truth contours and dashed red prediction contours, and separate 3D renderings illustrated representative blood-pool and hollow anatomy.

2.14 Exploratory Phase 2 case-characteristic analysis

For each Phase 2 case, CT dimensions, voxel spacing, field of view, ground-truth blood-pool volume, intensity statistics, and the number of empty predicted structures were extracted. Spearman rank correlations with case-level mean Dice were used descriptively to characterize failure patterns; these exploratory measures did not define exclusion rules or clinical thresholds.

Chapter 3
Results

3.1 Phase 1 configuration selection and validation performance

The learning-rate screen identified 0.01 as the selected configuration for CHD68. The primary checkpoint was epoch 1,250 with Dice plus cross-entropy loss and the 3d_fullres configuration. Extending training to 1,500 epochs did not materially increase validation Dice, consistent with convergence of the learning-rate runs. Across seven foreground structures, M0 achieved mean case-level Dice 0.846 ± 0.071 and median Dice 0.872 on the 14-case CHD68 validation set.

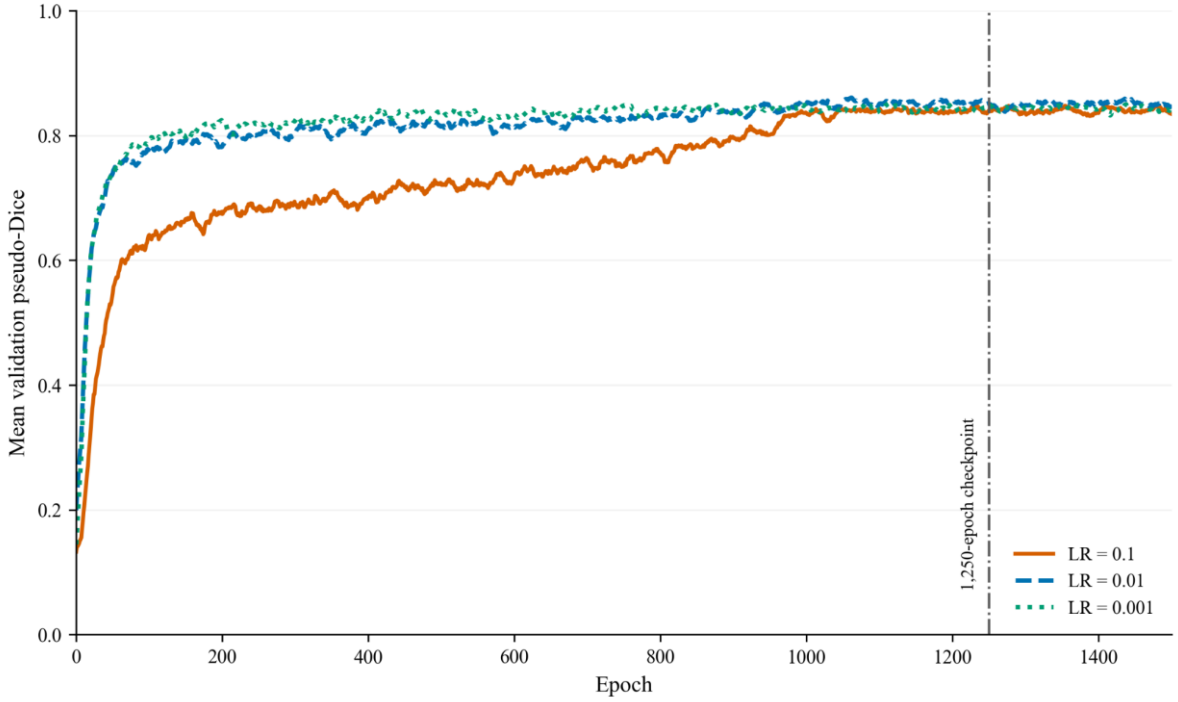


Figure 3.1. Phase 1 learning-rate screening.
Mean validation pseudo-Dice is shown across training epochs for learning rates 0.1, 0.01, and 0.001. The vertical reference line marks the 1,250-epoch checkpoint used for the primary experiments. Learning rate 0.01 was selected for subsequent phases.

Table 3.1. Selected Phase 1 performance at learning rate 0.01 and 1,250 epochs (n = 14 CHD68 validation cases).

Structure	Dice, mean ± SD	Median Dice	Mean IoU	Mean precision	Mean recall	HD95, median [IQR], mm
LV	0.863 ± 0.102	0.902	0.770	0.881	0.876	5.4 [3.9, 13.3]
RV	0.857 ± 0.096	0.891	0.759	0.843	0.884	4.9 [4.4, 10.9]
LA	0.889 ± 0.068	0.900	0.807	0.896	0.892	10.8 [3.7, 18.6]
RA	0.874 ± 0.048	0.884	0.780	0.893	0.868	21.9 [8.1, 38.4]
Myo	0.849 ± 0.115	0.892	0.751	0.868	0.837	6.6 [5.8, 7.9]
Ao	0.840 ± 0.111	0.867	0.737	0.852	0.858	28.1 [8.3, 30.6]
PA	0.750 ± 0.174	0.806	0.626	0.732	0.799	18.7 [6.7, 30.1]

LA had the highest mean Dice (0.889), followed by RA (0.874), LV (0.863), RV (0.857), myocardium (0.849), and Ao (0.840). PA had the lowest mean Dice (0.750) and the greatest Dice variability among the seven structures. HD95 was more variable than overlap metrics, with the largest median boundary errors observed for RA and PA.

3.2 Phase 2 performance on UCSF data

All 41 Phase 2 UCSF cases completed preprocessing and inference, with compatibility and geometry checks confirming valid inputs before evaluation. Across the six structures shared by CHD68 and UCSF, mean case-level Dice decreased from 0.845 ± 0.068 on CHD68 to 0.389 ± 0.316 on UCSF; the medians were 0.869 and 0.292. The mean difference (UCSF minus CHD68) was -0.457 (95% bootstrap CI, -0.555 to -0.353). The UCSF distribution was lower by two-sided Mann-Whitney testing ($p = 7.04 \times 10^{-6}$), with Cliff's delta -0.812.

3.3 Structure-level Phase 1 versus Phase 2 performance

Table 3.2. Structure-level Phase 1 versus Phase 2 Dice comparisons for the six shared blood-pool structures.

Structure	Phase 1 mean Dice	Phase 2 mean Dice	Change	Cliff's delta	Holm-adjusted p
LV	0.863	0.340	-0.523	-0.645	6.14×10^{-4}
RV	0.857	0.499	-0.357	-0.690	4.06×10^{-4}
LA	0.889	0.416	-0.474	-0.909	2.52×10^{-6}
RA	0.874	0.337	-0.537	-0.829	1.93×10^{-5}
Ao	0.840	0.379	-0.460	-0.714	2.96×10^{-4}
PA	0.750	0.361	-0.389	-0.620	6.14×10^{-4}

Note. Differences were computed from unrounded values.

Every shared structure had lower Dice in Phase 2 after Holm correction, and all structure-level Cliff's delta values were strongly negative. The largest mean decreases occurred for RA (-0.537), LV (-0.523), and LA (-0.474). RV had the highest Phase 2 mean Dice (0.499), while RA and LV had the lowest means (0.337 and 0.340, respectively).

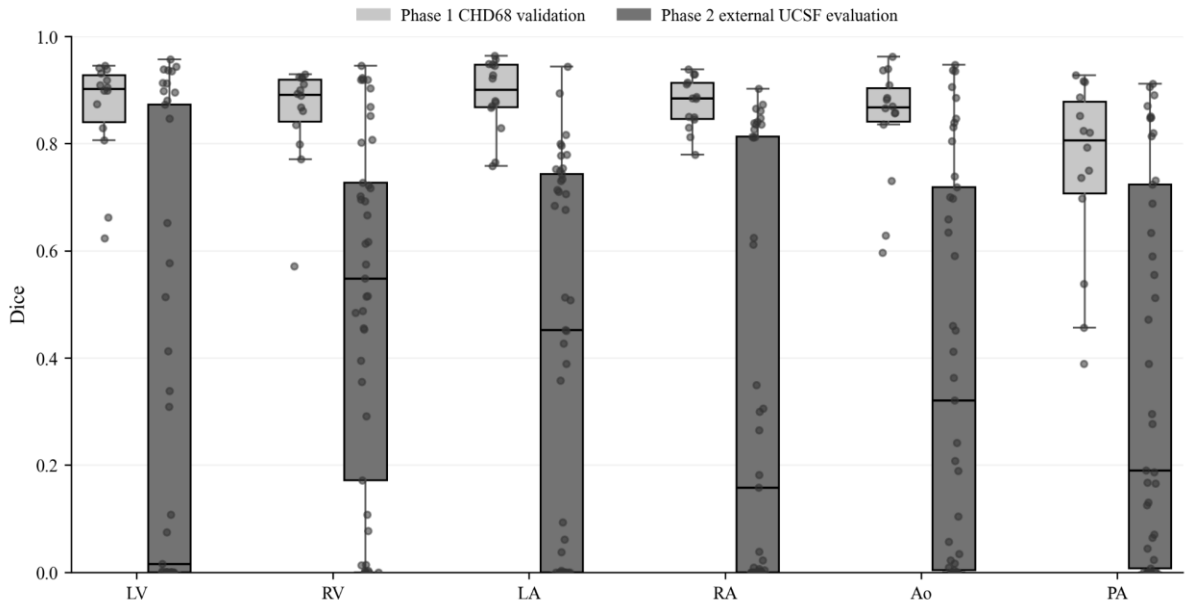


Figure 3.2. Phase 1 versus Phase 2 shared-structure Dice distributions.
Shared-structure Dice distributions are shown for Phase 1 CHD68 validation (n = 14) and Phase 2 UCSF evaluation (n = 41). Boxes show the median and interquartile range, and points show individual case-structure Dice values. All six UCSF distributions were lower and substantially broader.

3.4 Phase 2 failure patterns and case heterogeneity

Table 3.3. Phase 2 segmentation metrics and prediction failures (n = 41 UCSF cases).

Structure	Dice, median [IQR]	Mean precision*	Mean recall (all 41 cases)	HD95, median [IQR], mm*	Empty predictions
LV	0.016 [0.000, 0.872]	0.624	0.323	12.0 [3.7, 27.7]	13/41
RV	0.548 [0.171, 0.727]	0.560	0.555	14.4 [6.8, 20.6]	4/41
LA	0.452 [0.000, 0.743]	0.453	0.466	34.0 [23.5, 49.9]	3/41
RA	0.158 [0.000, 0.813]	0.569	0.353	36.7 [22.4, 46.4]	11/41
Ao	0.321 [0.004, 0.719]	0.597	0.354	50.2 [34.9, 64.4]	5/41
PA	0.190 [0.007, 0.723]	0.369	0.385	23.4 [16.7, 38.0]	0/41

*Precision and HD95 are undefined for empty predictions and are summarized over nonempty predictions only. Recall is summarized over all 41 cases; an empty prediction therefore contributes recall = 0. HD95 was calculated using the largest predicted connected component and the complete ground-truth structure.

Complete omissions occurred in 13 of 41 LV predictions, 11 RA predictions, five Ao predictions, four RV predictions, and three LA predictions; PA was present in every case. Because HD95 is undefined when a prediction is empty, boundary-error summaries alone did not capture the full severity of these failures.

Performance varied markedly across cases. Nine of 41 cases achieved mean Dice at or above 0.80, whereas 17 scored below 0.20 and 10 below 0.10. Exploratory analysis showed the strongest rank association between ground-truth blood-pool volume and case-level mean Dice (Spearman rho = 0.839). Field-of-view width and volume were also positively associated with Dice (rho = 0.677 and 0.659), while median blood-pool HU showed a weaker positive association (rho = 0.375). Empty predicted structures were negatively associated with Dice (rho = -0.669).

3.5 Phase 3 blood-pool comparison on nine UCSF validation cases

Table 3.4. Paired case-level comparisons on nine UCSF blood-pool validation cases.

Model pair	First model mean	Second model mean	Mean difference	Bootstrap 95% CI	Holm-adjusted p	Cases improved
M0 → M1	0.488	0.754	+0.266	[+0.102, +0.445]	0.0156	8/9
M0 → M2	0.488	0.755	+0.267	[+0.115, +0.434]	0.0117	9/9
M1 → M2	0.754	0.755	+0.001	[-0.039, +0.035]	0.977	5/9

On the same nine UCSF validation cases, mean Dice across the six blood-pool structures was 0.488 for M0, 0.754 for M1, and 0.755 for M2. Relative to M0, M1 increased mean Dice by 0.266 (95% bootstrap CI, 0.102 to 0.445; Holm-adjusted p = 0.0156), with improvement in eight of nine cases. M2 increased mean Dice by 0.267 (95% CI, 0.115 to 0.434; Holm-adjusted p = 0.0117), with improvement in all nine cases.

M1 and M2 showed very similar performance on the shared UCSF validation cases. The mean paired difference for M2 minus M1 was +0.001 (95% bootstrap CI, -0.039 to +0.035; Holm-adjusted p = 0.977); M2 was higher in five cases and M1 in four.

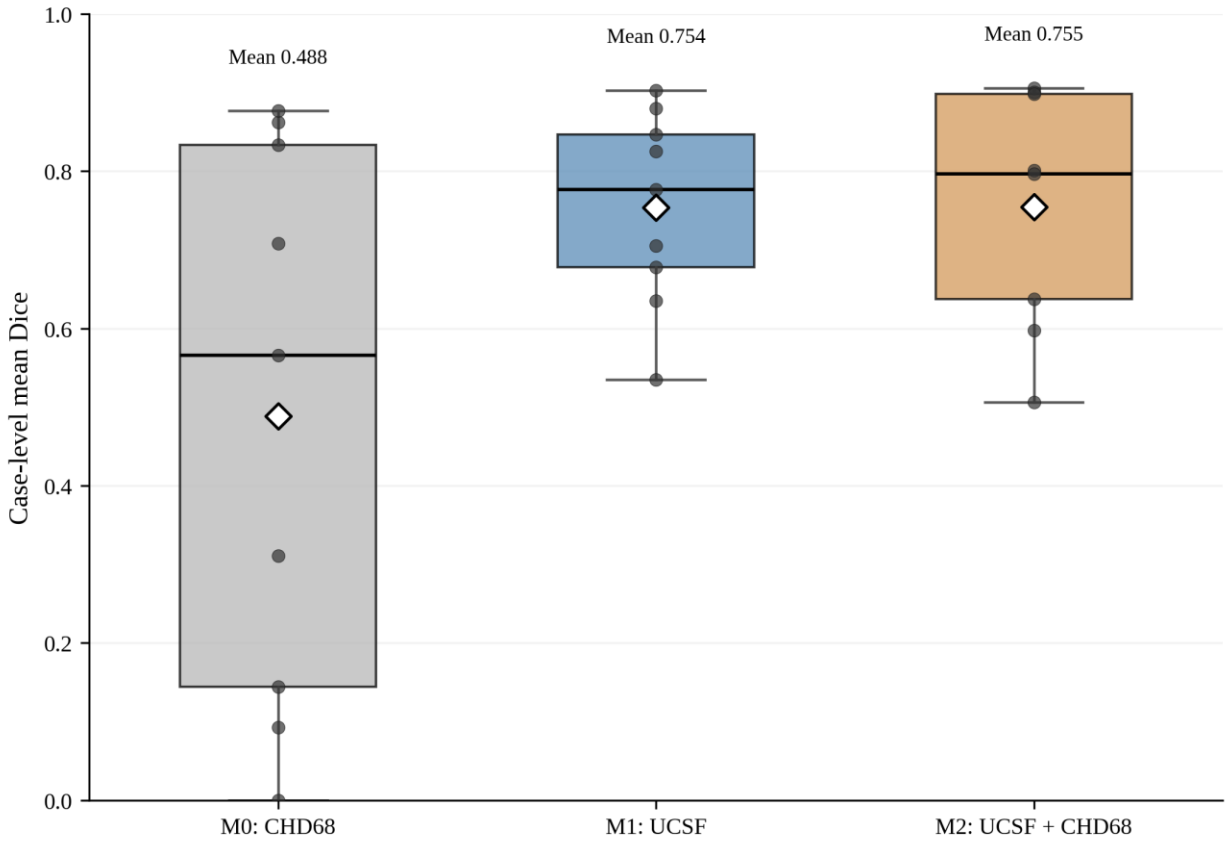


Figure 3.3. Phase 3 blood-pool comparison on nine UCSF validation cases.
M0, M1, and M2 are compared using case-level mean Dice across the six shared blood-pool structures. Dark circles show individual cases, boxes summarize each distribution, and white diamonds mark model means. Training with UCSF data increased performance substantially, while M1 and M2 showed similar performance.

3.6 Phase 3 blood-pool performance by structure

Table 3.5. Mean Dice by structure on the same nine UCSF validation cases.

Structure	M0	M1	M2
LV	0.472	0.761	0.777
RV	0.624	0.851	0.821
LA	0.469	0.628	0.644
RA	0.431	0.730	0.768
Ao	0.477	0.859	0.855
PA	0.458	0.694	0.664

Mean Dice was higher than M0 for all six shared structures in both UCSF-trained models. For M1, structure-level mean Dice ranged from 0.628 for LA to 0.859 for Ao; for M2, it ranged from 0.644 for LA to 0.855 for Ao. The similar structure-level results matched the case-level comparison.

3.7 M2 performance on CHD68 validation cases

M2 was also evaluated on the same 14 CHD68 validation cases used for the Phase 1 six-label analysis. Mean case-level Dice across the six structures was 0.836 for M2 compared with 0.845 for M0, a difference of approximately -0.009. The paired bootstrap 95% confidence interval for M2 minus M0 was -0.027 to +0.007, consistent with only a small average change on CHD68.

Table 3.6. CHD68 six-structure means for M0 and M2 on the same 14 validation cases.

Structure	M0 mean Dice	M2 mean Dice	Difference
LV	0.863	0.863	+0.001
RV	0.857	0.846	-0.011
LA	0.889	0.890	+0.001
RA	0.874	0.866	-0.008
Ao	0.840	0.831	-0.009
PA	0.750	0.722	-0.028

Note. Differences were computed from unrounded values.

3.8 Phase 3 ten-label hollow segmentation

M3 was evaluated on 14 held-out UCSF cases. Mean case-level Dice across ten hollow structures was 0.469 ± 0.160, with median 0.527. Structure-level mean Dice ranged from 0.320 for IVC to 0.583 for Ao. Ao (0.583), LA (0.574), and RA (0.539) had the highest means, whereas IVC (0.320), SVC (0.365), and coronary arteries (0.430) were among the lowest.

Table 3.7. M3 ten-label hollow performance by structure (n = 14 UCSF validation cases).

Structure	Mean Dice	Mean IoU	Mean precision	Mean recall	HD95, median [IQR], mm*	Empty predictions
LA	0.574	0.436	0.640	0.565	2.1 [1.0, 4.2]	1/14
LV	0.478	0.357	0.600	0.437	2.7 [1.2, 18.7]	1/14
RA	0.539	0.402	0.647	0.511	1.9 [1.1, 7.8]	1/14
RV	0.487	0.358	0.566	0.492	2.1 [1.3, 16.7]	1/14
Ao	0.583	0.435	0.635	0.600	19.0 [10.2, 29.1]	1/14
PA	0.464	0.324	0.518	0.472	13.4 [8.3, 23.0]	0/14
SVC	0.365	0.233	0.462	0.400	38.9 [27.7, 50.1]	1/14
IVC	0.320	0.215	0.336	0.384	16.9 [5.6, 31.3]	0/14
PV	0.452	0.304	0.485	0.498	43.9 [36.5, 84.5]	0/14
Coronaries	0.430	0.296	0.564	0.411	37.2 [31.9, 48.0]	1/14

*HD95 is undefined for empty predictions; median [IQR] values use nonmissing cases only. HD95 was calculated using the largest predicted connected component and the complete ground-truth structure.

Dice distributions were broad for several hollow structures. Seven empty structure predictions occurred in a single severe case-level failure; IVC, PA, and PV were predicted in all 14 cases.

Because hollow targets are thinner and more surface-like than blood-pool targets, their Dice values are particularly sensitive to wall thickness, branch extent, and truncation and should not be directly compared numerically with blood-pool Dice values.

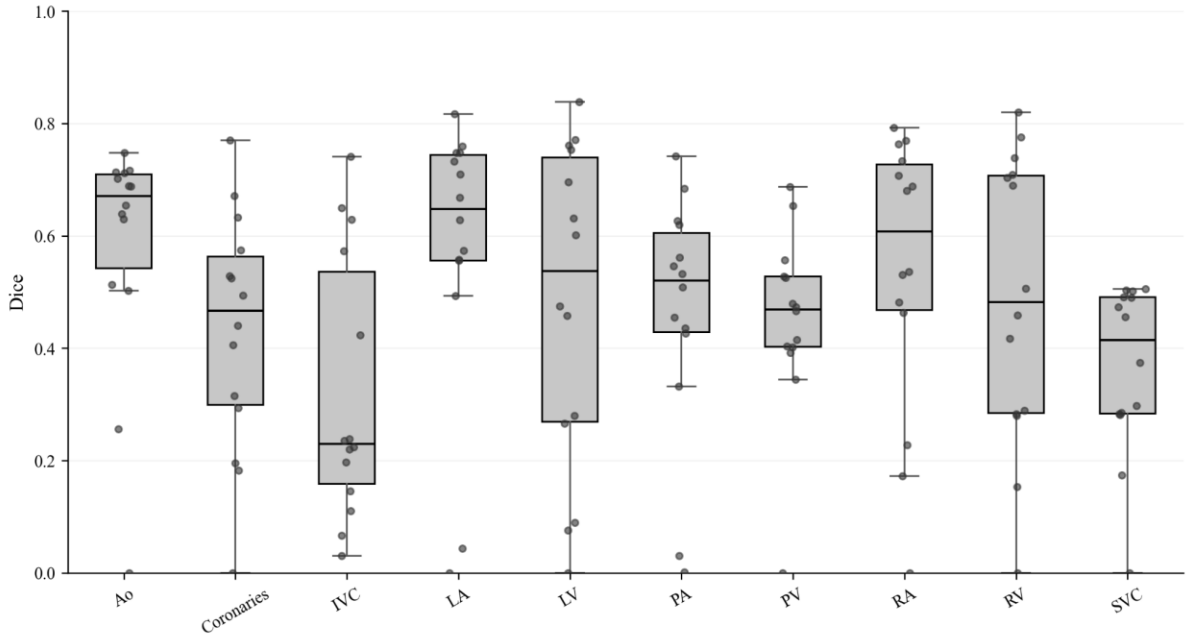


Figure 3.4. M3 hollow segmentation performance by structure.
Per-structure Dice distributions are shown for M3 on 14 held-out UCSF cases. Boxes show the median and interquartile range, and points show individual case-structure Dice values. Performance was heterogeneous across structures and cases, with greater variability for several smaller or branching vascular targets.

3.9 Ground-truth-bounded hollow analysis

The ground-truth-bounded analysis increased mean case-level Dice from 0.469 to 0.484 after prediction voxels outside the single ground-truth-defined whole-heart box were removed, an absolute increase of +0.015. The largest structure-level changes were SVC (+0.049), Ao (+0.039), IVC (+0.033), PV (+0.017), and PA (+0.010); chambers and coronary arteries changed little or not at all.

The modest overall increase shows that distal extent and truncation contributed to some vascular disagreement but did not explain most residual hollow overlap error. Because the allowable extent was defined from ground truth, this analysis is optimistic and does not represent a postprocessing step available during deployment.

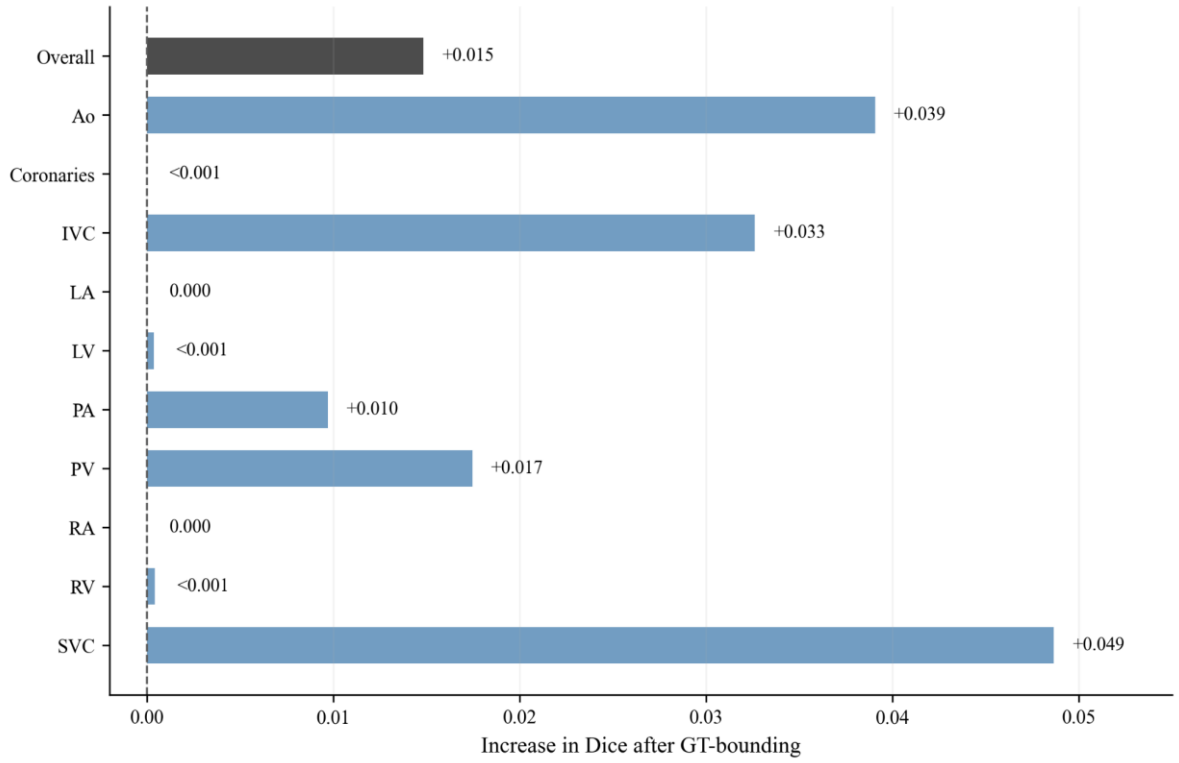


Figure 3.5. Change in M3 Dice after ground-truth bounding.
Bars show the change in mean Dice after prediction voxels outside a single whole-heart bounding box defined from the union of ground-truth labels were removed while labels inside the box were unchanged. The overall increase was +0.015, with the largest changes in SVC, Ao, IVC, and PV.

3.10 Qualitative evaluation of Phase 3 validation cases

Worst, median, and best cases by case-level mean Dice were selected for each Phase 3 model.

The blood-pool examples showed close overlap in higher-performing cases and increasing regional disagreement in lower-performing cases. M3 included one near-complete failure, but the median and best hollow examples preserved much of the central cardiac geometry while differing at surfaces, small vessels, and distal branch termination. In these examples, modest hollow Dice did not necessarily correspond to loss of the major 3D cardiac shape.

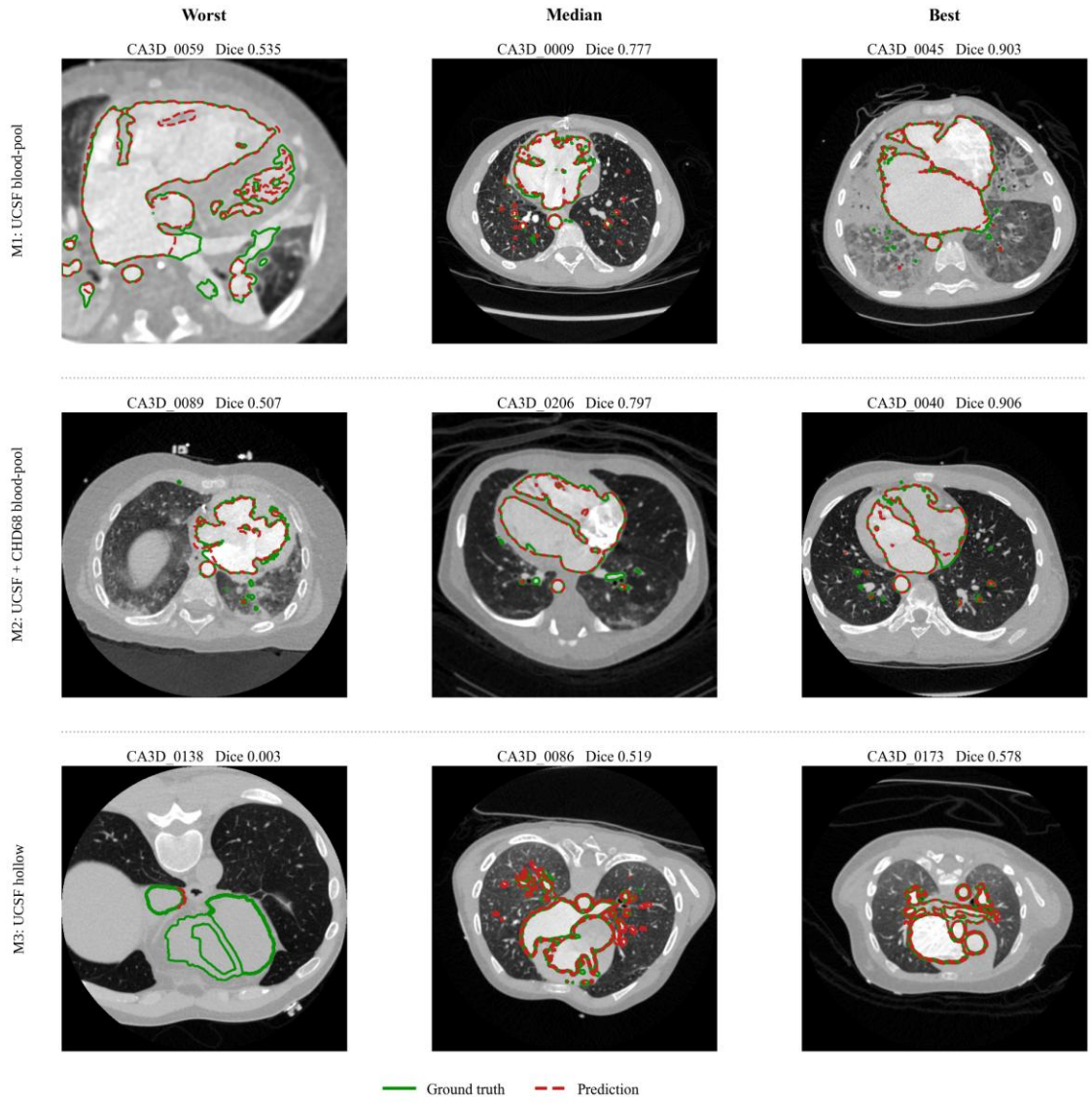


Figure 3.6. Representative 2D Phase 3 segmentation examples.

Worst, median, and best cases were selected by case-level mean Dice. Ground truth is shown with solid green contours and prediction with dashed red contours. For each case, the displayed axial CT slice contains the largest union of ground-truth labeled voxels.

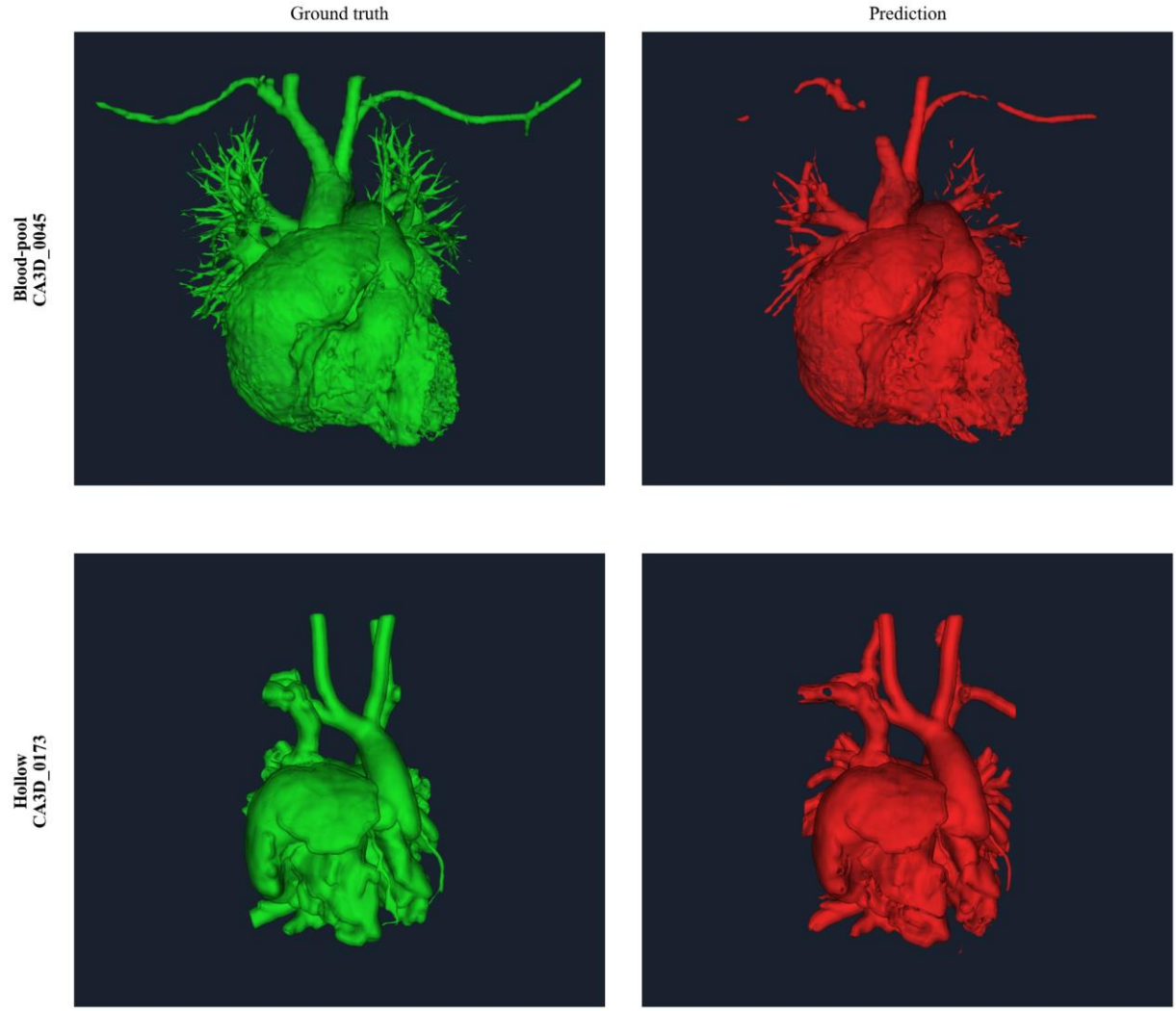


Figure 3.7. Representative 3D blood-pool and hollow segmentations.
The top row shows blood-pool ground truth and prediction for CA3D_0045. The bottom row shows hollow ground truth and the ground-truth-bounded M3 prediction for CA3D_0173. Green indicates ground truth and red indicates prediction. For the hollow example, the expert ground-truth extent defines clinically and 3D-printing-relevant truncation bounds, illustrating how distal extent can affect visual and quantitative agreement.

Chapter 4
Discussion

4.1 Principal findings and hypothesis resolution

The Phase 1, Phase 2, and Phase 3 blood-pool hypotheses were supported. Phase 1 exceeded the prespecified mean Dice threshold of 0.80 on held-out CHD68 cases (0.846). Phase 2 performance decreased substantially when M0 was applied to UCSF data (mean Dice 0.389). Phase 3 then showed that training with UCSF cases improved blood-pool performance on the same nine held-out UCSF cases: mean Dice increased from 0.488 for M0 to 0.754 for M1 and 0.755 for M2.

M2 did not improve UCSF performance beyond M1, but it largely retained performance on the CHD68 validation cases (0.836 versus 0.845 for M0). Together, these results indicate that exposure to UCSF training examples was important for improved UCSF performance, while adding CHD68 cases provided no measurable additional benefit on the small UCSF validation set.

The M3 hollow experiment was more mixed. Mean Dice was 0.469 across ten structures, with greater difficulty in several small or branching vessels. The ground-truth-bounded analysis increased mean Dice only to 0.484, indicating that truncation contributed to some vascular penalties but did not account for most error. At the same time, qualitative review showed that several nonfailure hollow predictions retained recognizable cardiac geometry despite modest overlap. This is important because the hollow target more closely matches the representation used for CA3D+ 3D modeling and printing and is not numerically equivalent to filled blood-pool segmentation.

4.2 Phase 1 configuration screening

Phase 1 established a stable configuration for the rest of the study. The learning-rate 0.01 run at 1,250 epochs achieved strong overlap across the four chambers, myocardium, and aorta, with mean Dice values between 0.840 and 0.889. Extending training to 1,500 epochs did not materially increase Dice, indicating that the learning-rate runs had reached convergence for the purposes of configuration selection.

Pulmonary artery segmentation was the most difficult Phase 1 target, with mean Dice 0.750 and greater variability than the larger chambers. Branching geometry and the choice of distal vessel extent can make PA boundaries more difficult to reproduce, illustrating why whole-structure Dice should be interpreted together with boundary error and anatomical review.

4.3 Performance on UCSF data and cross-dataset generalization

Phase 2 showed that strong validation performance on CHD68 did not guarantee similar performance on UCSF cases. The decrease from 0.845 to 0.389 across the six shared structures was large but highly heterogeneous: nine UCSF cases retained mean Dice at or above 0.80, whereas 17 scored below 0.20 and several approached complete prediction collapse. The wide range of UCSF performance showed that the decrease was highly case dependent rather than uniform.

Technical checks made a single global implementation error unlikely: CHD68 validation performance remained high, converted CT-mask geometry was verified, and some UCSF cases achieved high Dice after preprocessing.

The reduction in Phase 2 performance could reflect multiple differences between CHD68 and UCSF, including patient scale, congenital and postoperative anatomy, acquisition characteristics, contrast timing, field of view, intensity representation, and annotation conventions. Even

identically named structures may be represented differently because of connected-vessel extent, surgically altered pathways, or other annotation choices. Exploratory analysis also associated lower Dice with smaller ground-truth blood-pool volume and smaller field of view, although these measures were correlated and cannot identify a single causal factor. Phase 2 therefore supports the broader point that evaluation on clinically representative local data is necessary before relying on a model developed elsewhere.

4.4 UCSF-trained blood-pool models

Evaluation on the same nine UCSF validation cases showed that training with UCSF examples substantially improved blood-pool performance. M1 increased mean Dice by 0.266 relative to M0 and improved eight of nine cases; M2 increased mean Dice by 0.267 and improved all nine cases. Mean Dice increased across all six shared structures rather than being driven by a single label.

M1 and M2 showed very similar performance on UCSF validation: their means differed by 0.001, and the paired bootstrap 95% confidence interval for M2 minus M1 (-0.039 to +0.035) included zero. Adding CHD68 cases therefore provided no measurable additional UCSF benefit in this small validation set. M2 nevertheless retained high CHD68 performance, with mean six-structure Dice 0.836 compared with 0.845 for M0.

These findings indicate that at least part of the Phase 2 performance loss was recoverable through training with UCSF examples. At the same time, the small training and validation cohorts limit conclusions about whether M1 or M2 would be preferable for broader deployment.

4.5 Hollow segmentation and 3D-modeling relevance

The ten-label M3 experiment addressed the target most closely aligned with the CA3D+ modeling workflow. Hollow segmentations are thin, surface-like representations and require

explicit choices about wall thickness, distal vessel extent, branch inclusion, and junction boundaries. These properties make Dice particularly sensitive to small surface shifts and truncation differences. Blood-pool and hollow Dice values therefore should not be directly compared as equivalent measures of model quality.

Mean M3 Dice was 0.469, with the greatest variability in several smaller or branching vessels. The ground-truth-bounded analysis increased mean Dice by only +0.015 overall but produced larger changes for SVC, Ao, IVC, and PV, supporting the observation that distal extent contributed to some vascular penalties. Because most error remained after bounding, annotation extent alone does not explain the lower hollow overlap.

Qualitative review adds an important perspective. The median and best hollow examples preserved much of the central cardiac shape despite modest Dice, suggesting that some predictions could still provide useful initialization for expert correction, mesh generation, and 3D printing. This thesis did not measure correction time, printability, or surgeon acceptability, so clinical utility has not yet been demonstrated. The current M3 model should therefore be viewed as an initial feasibility result for a distinct segmentation target that requires further refinement rather than as an autonomous clinical segmentation system.

4.6 Annotation consistency and workflow implications

The results also emphasize that a segmentation label is defined by the annotation protocol as well as by anatomy. Differences in label exclusivity, vessel termination, surgical pathways, wall thickness, and hollow versus filled representations change both what a model learns and how its predictions are scored. Consistent label names are therefore not enough when datasets were produced under different segmentation rules.

For CA3D+, an expert-in-the-loop workflow remains the appropriate goal. An automated system could provide initial masks or surfaces that a 3D engineer or clinician reviews and corrects before downstream modeling. The clinically meaningful endpoint is therefore not Dice alone: useful automation must preserve important anatomy, avoid major omissions, remain straightforward to edit, and reduce the time required to create an acceptable final model.

4.7 Limitations

Several limitations should be considered when interpreting these findings. Performance estimates for the trained models were based on relatively small held-out validation cohorts and single validation splits and may therefore be sensitive to cohort composition. Because the same 14 CHD68 validation cases were used for both configuration selection and performance reporting, Phase 1 performance was not evaluated on a separate unseen test cohort. Cross-validation across multiple splits would provide more stable estimates of model performance and configuration selection. Phase 2 included 41 UCSF cases, while the Phase 3 blood-pool comparison used nine held-out UCSF cases and M3 used 14, limiting the precision and generalizability of these estimates.

The reduction in performance from CHD68 to UCSF could reflect several factors acting simultaneously, including differences in patient scale, congenital and postoperative anatomy, acquisition characteristics, contrast timing, field of view, intensity representation, and annotation conventions. This study demonstrated a substantial cross-dataset performance decrease but was not designed to isolate the contribution of any single factor.

The Phase 2 case-characteristic analysis was exploratory, and ground-truth blood-pool volume is explanatory rather than a pre-segmentation predictor. Precision and HD95 were undefined for empty predictions, so summaries over nonempty cases are necessarily optimistic. Phase 3 models

were evaluated only on their designated held-out splits, and the UCSF-trained models were not evaluated on an external institutional cohort.

The hollow experiment also lacked a formal interobserver annotation study, so the contribution of annotation variability to measured error is unknown; qualitative examples were selected by Dice rather than a blinded acceptability framework and were intended only as illustrations. Most importantly, this thesis evaluated segmentation performance rather than prospective clinical benefit. It did not measure expert correction time, mesh-repair burden, print success, surgical decision-making, or patient outcomes. The ground-truth-bounded analysis was intentionally optimistic because the bounding box was defined from the reference annotation.

4.8 Future directions

The first priority is refinement and standardization of the hollow training targets. Annotation rules should define clinically appropriate wall thickness, vessel extent, branch inclusion, junction boundaries, and truncation planes. Cases with large distal extent disagreement should be re-reviewed so that the model is trained against consistent endpoints rather than annotation variability.

The refined training set should then be expanded beyond the current subset to use the broader CA3D+ archive of more than 200 cases. Additional examples should deliberately include infants, uncommon anatomies, postoperative pathways, smaller structures, and underrepresented acquisition conditions. Increasing both the number and diversity of well-standardized cases is likely to be more valuable than changing architectures before the target itself is consistent.

After retraining on refined hollow targets, evaluation should retain Dice and surface metrics but add endpoints that reflect the intended use: expert correction time, mesh-repair burden, anatomical acceptability, printability, and whether automated initialization shortens the complete

CA3D+ workflow. These measures would determine whether anatomically coherent predictions with imperfect overlap actually reduce manual work.

Subsequent development should also include automated failure checks, such as missing-structure detection, implausible predicted volumes, model disagreement, or case-level quality scores, so problematic outputs are flagged before downstream modeling. Broader validation should then use additional UCSF cases and, where possible, data from another institution to test whether the refined blood-pool and hollow models remain reliable outside the current splits.

Together, these steps would determine whether automated segmentation can move beyond overlap-based performance and meaningfully reduce the expert effort required for patient-specific cardiac 3D modeling. The present work establishes a foundation for that development by demonstrating strong UCSF blood-pool performance and identifying the key challenges that remain for hollow segmentation.

References

1. van der Linde D, Konings EEM, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJM, Roos-Hesselink JW. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J Am Coll Cardiol. 2011;58(21):2241-2247. doi:10.1016/j.jacc.2011.08.025.
2. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z.
3. Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, Jäger PF. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2024. Lecture Notes in Computer Science. Vol 15009. Springer; 2024:488-498. doi:10.1007/978-3-031-72114-4_47.
4. Alnasser TN, Abdulaal L, Maiter A, Sharkey M, Dwivedi K, Salehi M, Garg P, Swift AJ, Alabed S. Advancements in cardiac structures segmentation: a comprehensive systematic review of deep learning in CT imaging. Front Cardiovasc Med. 2024;11:1323461. doi:10.3389/fcvm.2024.1323461.
5. Xu X, Wang T, Shi Y, Yuan H, Jia Q, Huang M, Zhuang J. Whole Heart and Great Vessel Segmentation in Congenital Heart Disease Using Deep Neural Networks and Graph Matching. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2019. Lecture Notes in Computer Science. Vol 11765. Springer; 2019:477-485. doi:10.1007/978-3-030-32245-8_53.

6. Yao Z, Xie W, Zhang J, Yuan H, Huang M, Shi Y, Xu X, Zhuang J. Graph matching and deep neural networks based whole heart and great vessel segmentation in congenital heart disease. Sci Rep. 2023;13:7558. doi:10.1038/s41598-023-34013-1.

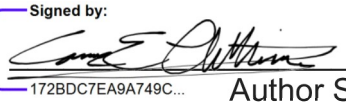
7. Zhu Y, Li H, Cao B, Huang K, Liu J. A novel hybrid layer-based encoder-decoder framework for 3D segmentation in congenital heart disease. Sci Rep. 2025;15:11891. doi:10.1038/s41598-025-96251-9.

8. Liang B, Peng F, Luo D, Zeng Q, Wen H, Zheng B, et al. Automatic segmentation of 15 critical anatomical labels and measurements of cardiac axis and cardiothoracic ratio in fetal four chambers using nnU-NetV2. BMC Med Inform Decis Mak. 2024;24:128. doi:10.1186/s12911-024-02527-x.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

Signed by:

172BDC7EA9A749C...
Author Signature

08/24/2026
Date