Multi-task learning for joint weakly-supervised segmentation and aortic arch anomaly classification in fetal cardiac MRI

Paula Ramirez https://orcid.org/0000-0002-0705-5296   paula.ramirez_gilliland@kcl.ac.uk

Alena Uus https://orcid.org/0000-0001-5796-2145   alena.uus@kcl.ac.uk

Milou P.M. van Poppel https://orcid.org/0000-0002-1739-4726   milou.van_poppel@kcl.ac.uk

Irina Grigorescu https://orcid.org/0000-0002-9756-3787   irina.grigorescu@kcl.ac.uk

Johannes K. Steinweg https://orcid.org/0000-0002-3366-0932   johannes.steinweg@kcl.ac.uk

David F.A. Lloyd https://orcid.org/0000-0003-1759-6106   david.lloyd@kcl.ac.uk

Kuberan Pushparajah https://orcid.org/0000-0003-1541-1155   kuberan.pushparajah@kcl.ac.uk

Andrew P. King https://orcid.org/0000-0002-9965-7015   andrew.king@kcl.ac.uk

Maria Deprez https://orcid.org/0000-0002-2799-6077   maria.deprez@kcl.ac.uk

School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK.

Abstract

Congenital Heart Disease (CHD) is a group of cardiac malformations present already during fetal life, representing the prevailing category of birth defects globally. Our aim in this study is to aid 3D fetal vessel topology visualisation in aortic arch anomalies, a group which encompasses a range of conditions with significant anatomical heterogeneity. We present a multi-task framework for automated multi-class fetal vessel segmentation from 3D black blood T2w MRI and anomaly classification. Our training data consists of binary manual segmentation masks of the cardiac vessels' region in individual subjects and fully-labelled anomaly-specific population atlases. Our framework combines deep learning label propagation using VoxelMorph with 3D Attention U-Net segmentation and DenseNet121 anomaly classification. We target 11 cardiac vessels and three distinct aortic arch anomalies, including double aortic arch, right aortic arch, and suspected coarctation of the aorta. We incorporate an anomaly classifier into our segmentation pipeline, delivering a multi-task framework with the primary motivation of correcting topological inaccuracies of the segmentation. The hypothesis is that the multi-task approach will encourage the segmenter network to learn anomaly-specific features. As a secondary motivation, an automated diagnosis tool may have the potential to enhance diagnostic confidence in a decision support setting. Our results showcase that our proposed training strategy significantly outperforms label propagation and a network trained exclusively on propagated labels. Our classifier outperforms a classifier trained exclusively on T2w volume images, with an average balanced accuracy of 0.99 (0.01) after joint training. Adding a classifier improves the anatomical and topological accuracy of all correctly classified double aortic arch subjects.

Our code is available at
https://github.com/SVRTK/MASC-multi-task-segmentation-and-classification.

Keywords: Deep Learning, Fetal Cardiac Imaging, Congenital Heart Disease, Automated Diagnosis, Fetal Cardiac MRI, Aortic Arch Segmentation, Multi-Task Learning, Multi-Class Vessel Segmentation, Anomaly Segmentation

1. Introduction

Congenital heart disease (CHD) is the leading cause of mortality related to congenital defects (Mendis et al. (2011)). Accurate CHD diagnosis before birth is essential to inform appropriate early postnatal management, which is known to lead to improved patient outcomes both in terms of mortality and long-term morbidity (Brown et al. (2006); Mazwi et al. (2013)).

We present a fully automated weakly-supervised multi-task tool for multi-class fetal cardiac vessel segmentation and anomaly classification in 3D T2w MRI. Our intention is to facilitate fetal cardiac vessel visualisation for prenatal diagnostic reporting purposes, and provide the groundwork for automated vessel biometry and detection. We target three aortic arch anomalies: Right Aortic Arch (RAA), Double Aortic Arch (DAA), and suspected Coarctation of the Aorta (CoA).

In current clinical practice, vessel segmentation in fetal cardiac MRI is a time-consuming process based on thresholding followed by manual correction, resulting in a binary mask of fetal heart and vessels. An expert fetal cardiac clinician generally takes 1-2 hrs to complete the binary mask, which is a significant hurdle to wider translation outside the research setting. In addition, more refined multi-label segmentation would provide clinicians with better visualisation of individual vessels, and facilitate automated quantitative analysis, thus reducing reporting time and potentially improving the prediction of outcomes. However, manual multi-label segmentation would bring extra clinical burden and is therefore not currently performed in clinical practice. This is partly due to the image quality available, paired with the small fetal vessel size (often only 1-2 voxels wide).

Our proposed weakly supervised deep learning framework is able to leverage existing binary manual segmentations in conjunction with a small number of condition-specific multi-label atlases, to provide a fully automated and accurate multi-label segmentation of individual cardiac vessels. The technique is adaptable to clinical workflows as it does not require any manual input at inference time. We incorporate an aortic arch anomaly classifier into our framework, with the intention of both providing a clinically useful diagnostic tool and improving segmentation performance.

1.1 Imaging the fetal heart

2D Fetal echocardiography is commonly used for detecting CHD before birth, as it offers clear discernment of the cardiac chambers and cardiac functional measurements using Doppler flow. Extracting vessel positional and topological information from 2D echocardiography images alone is an extremely challenging task, requiring highly trained experts. Although 3D STIC (DeVore et al., 2003) is available in many clinical settings, the corruption by fetal motion renders this imaging modality less viable for clinical assessments.

Recently, fetal cardiac MRI has shown the potential as an adjunct to echocardiography in the detection of CHD prenatally (Lloyd et al., 2019), and its use is becoming widespread (Dong and Zhu, 2018; Salehi et al., 2021; Dong et al., 2020). State-of-the-art motion correction algorithms (Uus et al., 2020) address fetal CMR motion challenges, allowing the generation of high-quality 3D MRI of fetal cardiac vessels. T2w black blood 3D MRI offers excellent vascular visualisation (Lloyd et al., 2019), as the vessels appear dark in contrast with surrounding tissue such as the lungs. This is the principal motivation behind the choice

of imaging modality used in the present work, given the aim is segmentation of vascular structures in aortic arch anomalies.

1.2 Deep learning segmentation

In order to provide a robust assessment of the cardiovascular anatomy in CMR, anatomical structures require segmentation, i.e. voxelwise classification into an anatomical class. Deep learning has become widespread for automated segmentation (Hesamian et al., 2019) and adult CMR segmentation (Chen et al., 2020).

U-Net-based architectures (Ronneberger et al., 2015; Isensee et al., 2018) have been successfully employed for fetal brain and thorax MRI segmentation (Salehi et al., 2018; Uus et al., 2021). The fetal brain is the most commonly targeted fetal organ for automated segmentation in recent works (Keraudren et al., 2014; Khalili et al., 2019; Ebner et al., 2020; Payette et al., 2020; Salehi et al., 2018), due to the growing interest in early brain development. In addition, as a rigid, heterogeneous organ constrained with the fetal skull the brain represents an attractive target for these methods, compared to the fetal heart which regularly deforms with the rapid fetal cardiac cycle. Most forms of CHD also present significant anatomical variation in terms of vessel sizes and relative positions, making the application of the methods particularly challenging.

Deep learning has been extensively applied to adult CMR segmentation in CHD (Arafati et al., 2019). Yu et al. (2016) present a 3D fractal network for whole heart and great vessel segmentation, while Rezaei et al. (2018) propose a framework comprising a three-stage cascade of conditional GANs. Xu et al. (2019) address the anatomical variability challenge in CHD by employing deep neural networks to segment the myocardial blood pool and chambers, coupled with graph matching for vessel identification. Fetal cardiac datasets, however, differ significantly from those acquired in postnatal life. For example, motion corruption is particularly common in fetal imaging, due to gross fetal, fetal cardiac and maternal respiratory motion. The fetal cardiac structures are also diminutive (most vessels are only millimetres in diameter), and are subject to low signal-to-noise ratio (SNR) and resolution with no option for intravenous contrast-enhancing agents. There is also differing baseline contrast due to the heart being surrounded by fluid filled lungs, as well as issues around temporal resolution due to the rapid fetal cardiac cycle. The technical approaches required for fetal cardiac imaging, such as fast acquisition sequences (Patel et al., 1997; Semelka et al., 1996) and high quality slice-to-volume registration algorithms (Kuklisova-Murgasova et al., 2012; Uus et al., 2020) are therefore highly specialised for this application, and generate unique output data which are not comparable to postnatal imaging methods. A bespoke approach for the application of deep learning methods is therefore required.

1.3 Atlas-guided segmentation

Our training dataset consists of partially labelled subjects (binary manually segmented labels) and fully-labelled atlases (Fig. 1). This type of setup is not uncommon in the medical imaging field, with exhaustive works addressing this challenge (Peng and Wang, 2021).

Atlas-based label propagation has been widely used for medical image segmentation (Aljabar et al., 2009; Heckemann et al., 2006). It uses image registration to transfer labelling

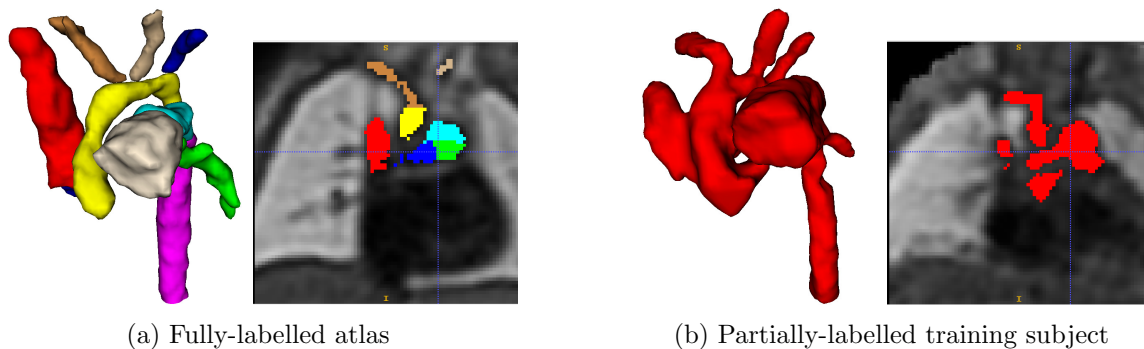(a) Fully-labelled atlas (b) Partially-labelled training subject

Figure 1: Illustration of our dataset setup.

information from a given atlas to individual subjects. We propose to use **VoxelMorph** (Balakrishnan et al., 2019) for propagating multi-class labels from anomaly-specific atlases to fetal subjects.

Deep learning atlas-based segmentation approaches include Dinsdale et al. (2019), which presents a binary mask warping framework based on spatial transformer networks; Xu and Niethammer (2019) propose to jointly train registration and segmentation networks to tackle partially labelled data. Similarly, in Sinclair et al. (2022), registration and segmentation are performed within the same framework, with the notable addition of a population-derived atlas being constructed in the process. Alternative strategies include one-shot or few-shot segmentation techniques (Zhao et al., 2019), where synthetic labelled data is generated using VoxelMorph and used to deal with partially labelled datasets.

However, label propagation strategies are always limited by registration accuracy, hence in this work we propose to combine the benefits of atlas-based and deep learning based segmentation and use the atlas-propagated labels to train a convolutional neural network (CNN) for the task of fetal heart segmentation.

1.4 Anomaly classifier

We also propose to leverage the notable topological distinction between aortic arch anomalies (Fig. 2) to explore the addition of an anomaly classifier into our framework, based on the pipeline presented by Puyol-Antón et al. (2021) which included a race classifier in combination with a segmentation network. A similar approach has been taken by Mehta et al. (2018), who address joint segmentation and classification of breast biopsy images, with a U-Net-based network predicting both a standard segmentation and discriminative map which are combined for tissue classification.

The benefits of joint classifier and segmenter network training have also been evidenced in the field of computer vision. Liao et al. (2016) propose an approach which in essence is the reverse of our strategy, namely, features from a classifier network are used to train a scene segmentation network. MultiNet (Teichmann et al. (2018)) also follows an analogous architecture to ours, combining a shared encoder with a task-specific decoder.

Our primary motivation is to explore whether the addition of a classifier can improve segmentation performance, in particular ensuring the correct topology of the automated

segmentation outputs. As a secondary motivation, we explore the possibility automated diagnosis of aortic arch anomalies as a potential aid to clinical decision making.

1.5 Contributions

In this work, we present a **weakly supervised multi-task framework for automated multi-class fetal cardiac vessel segmentation and anomaly classification** from 3D reconstructions of T2w black blood MRI images. We propose a deep learning framework, addressing three aortic arch anomalies.

We expand on our prior research (Ramirez Gilliland et al., 2022) by incorporation of an anomaly classifier which improves segmentation discernment between anomalies. We also include detailed ablation studies to validate individual elements of our multi-task framework, which now consists of training using labels propagated from the anomaly-specific atlases and manual labels in individual images, while simultaneously classifying the anomaly from the predicted segmentation. This way we can leverage valuable anomaly-specific features learnt from a classifier to improve our segmentation output. A key novelty is the application of our framework to fetal Cardiac Magnetic Resonance (CMR) data, which is to date largely unexplored. Our framework expands on our previous work (Uus et al., 2022b), which for the first time proposed an automated segmentation of fetal cardiac vessels by training a CNN on labels propagated from anomaly-specific atlases.

Our clinical contribution is to aid 3D vessel topology visualisation for clinical reporting and training purposes and to increase confidence in diagnostic accuracy. We propose a multi-class approach to enable isolated visualisation of the anomaly area and affected vessels. Our approach does not require a fully-labelled training set, only subject-specific binary labels alongside multi-class anomaly-specific atlases; and thus is adaptable to clinical environments. These labels are only required for training, segmentation during inference is carried out on fully unlabelled data, from a single network.

2. Methods

2.1 Dataset description

2.1.1 AORTIC ARCH ANOMALIES

Our automated segmentation tool is aimed at subjects with Right Aortic Arch (RAA) with aberrant left subclavian artery (ALSA), Double Aortic Arch (DAA), and suspected Coarctation of the Aorta (CoA). These three distinct anomalies are depicted in Fig. 2.

Briefly, in CoA, there is a narrowing of the distal aortic arch which can require urgent postnatal intervention if severe. In RAA the aortic arch passes to the right of the trachea (as opposed to the normal position on the left) with an associated risk of airway and/or oesophageal compression, particularly when associated with an ALSA. In DAA, both right and left aortic arches are present, forming a "vascular ring" around the trachea and oesophagus, which in turn is associated with a higher risk of postnatal symptoms requiring surgical intervention. Whilst each of these categories presents a distinct topological phenotype, there is significant heterogeneity within each group in terms of vessel size and morphology.

Our dataset is representative of patients that are referred to a specialist fetal cardiac service, and therefore no controls are included in this study. Nonetheless, cases with Suspicion of CoA have the same vessel topology as healthy controls.
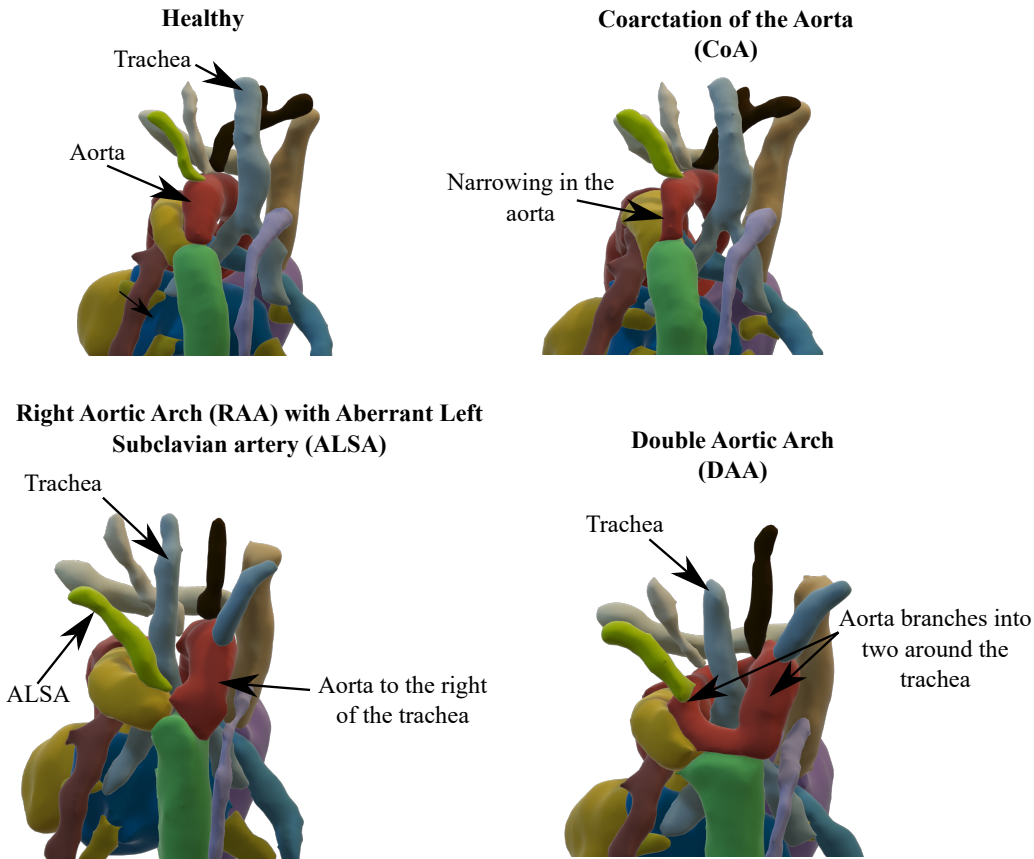

Figure 2: Aortic arch anomaly illustration (atlases by Uus et al. (2022b)).

2.1.2 DATA SPECIFICATIONS

We employ 3D reconstructions of T2w black blood MRI. Our dataset consists of 195 fetal subjects with suspected coarctation of the aorta (CoA, N=94), Right Aortic Arch (RAA) with ALSA (N=72) and Double Aortic Arch (DAA, N=29), 31.4±1.5 weeks mean gestational age (min=29 weeks, max=36 weeks). In addition to the primary diagnosis of suspected CoA, RAA, or DAA, 56 cases presented secondary diagnoses affecting additional anatomical areas, outside of the segmentation area (aside from RAA>DAA). These are included in Table 1.

The datasets were acquired at Evelina London Children's Hospital using a 1.5 Tesla Ingenia MRI system and a T2-weighted SSFSE sequence (RT=20,000 ms, ET=50ms, FA=90°, voxel size=1.25 × 1.25 mm, slice thickness=2.5 mm and slice overlap=1.25 mm). All research participants provided written informed consent. The raw datasets consisted of six to 12 multi-slice stacks of 2D images, covering the fetal thorax in three orthogonal planes.
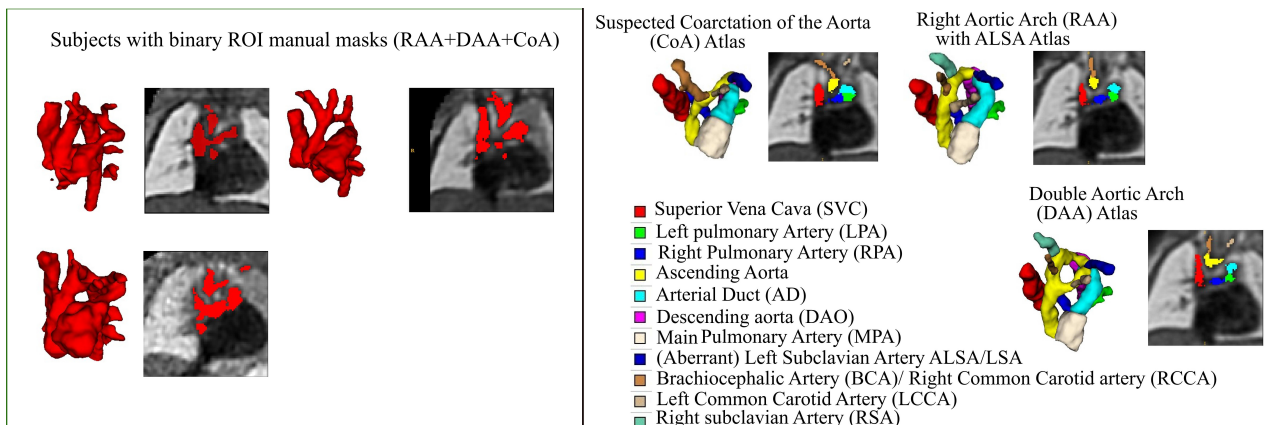
(a) 3D reconstructed T2w MRI images with binary manually segmented labels.

(b) 3D T2w MRI aortic arch anomaly atlases with multi-class vessel segmentations

Figure 3: Depiction of our two dataset groups: individual fetal subjects with binary masks (3a), and multi-class anomaly-specific atlases (3b)

Table 1: Number of cases with an additional diagnosis. AVSD: atrioventricular septal defect, VSD: ventricular septal defect, MV: mitral valve, PAPVD: partial anomalous pulmonary venous drainage, LV/RV: left/right ventricle, AS: aortic valve stenosis, DV: ductus venosus, BAV: bicuspid aortic valve, R/LAD: right/left arterial duct, R/LAA: right/left aortic arch, ?=suspected.

Primary diagnosis	Additional diagnosis	Number of cases
Suspected CoA	(unbalanced) (A)VSD	17
	?Parachute MV	1
	Small pericardial effusion	1
	PAPVD	1
	LV < RV	1
	(aneurysmal) AS	3
	Aberrant DV	1
	Aneurysm	2
	(?)BAV	3
RAA with ALSA	(?) VSD	4
	RAD	1
DAA	RAA>LAA	10
	LAD	9
	Retro-aortic innominate vein	1
	VSD	1

We use images reconstructed both with Slice-to-Volume Registration (SVR) (Kuklisova-Murgasova et al., 2012; Kainz et al., 2015) and Deformable SVR (Uus et al., 2020, 2022a) (DSVR, higher quality, N=49) to 0.75 mm isotropic resolution, to ensure a varied dataset.

Trained clinicians manually segmented a binary vessels label for the majority of our subjects (N=181), Fig. 3a. We employ propagated atlas labels exclusively for the remaining unsegmented cases (N=14, Sec. 2.2). In order to achieve our multi-class output, we employ three fully-labelled atlases[1] (Uus et al. (2022b), see Fig. 3b), one per anomaly (RAA, DAA and CoA). These include 11 manually segmented vascular regions for RAA and DAA, and 10 for CoA cases.

We affinely register all subject images to the pertinent anomaly-specific atlas prior to training via MIRTK[2] (Schnabel et al., 2001; Rueckert et al., 1999b,a), and crop to a standardised cardiac vessels region after affine alignment (see Fig. 3). We split the subjects into a training set (N_{CoA}=71, N_{RAA}=54, N_{DAA}=21), validation set (N_{CoA}=3, N_{RAA}=3, N_{DAA}=3), and testing set (N_{CoA}=20, N_{RAA}=15, N_{DAA}=5). We normalise and rescale the intensities to between 0 and 1, and use a weighted random sampler to correct for the class imbalance problem.

2.2 Multi-task segmentation framework

Our benchmark framework is based on Ramirez Gilliland et al. (2022), where Attention U-Net (Oktay et al., 2018) is trained using both manual binary labels and multi-class labels propagated from an atlas (Uus et al., 2022b).

The input to the network is an MRI image, and the output is a multi-class segmentation (see Fig. 4). We use VoxelMorph (Balakrishnan et al., 2019) for label propagation, defining our atlases as moving images (m), and fetal subject images as fixed images (f).

Subsequent to VoxelMorph training, we generate our propagated labels and use these to train a CNN for segmentation. We keep the weights of VoxelMorph label propagation frozen while training our segmentation network. We train our proposed network with two losses: (1) a multi-class loss between the propagated labels and the network output; (2) a binary loss between the multi-class network output joined into a binary segmentation and the manual binary labels.

Finally, we attach a DenseNet121 (Huang et al., 2017) classifier to our softmax segmentation output, which predicts one out of three possible diagnoses (CoA, RAA or DAA). We train Attention U-Net and DenseNet121 jointly. This classifier has the option of incorporating the image as an extra channel, to make the classification less dependent on accurate segmentation.

2.3 Label propagation

We use VoxelMorph (Balakrishnan et al., 2019) for label propagation. We warp the labels from the pertinent anomaly-specific atlas (moving images, m) into each subject space (fixed images f), using the prior anomaly diagnosis knowledge to select the relevant atlas. Note that at testing or inference time, no prior diagnosis knowledge is required, and our framework offers an automated diagnosis prediction.

---

1. https://gin.g-node.org/SVRTK/
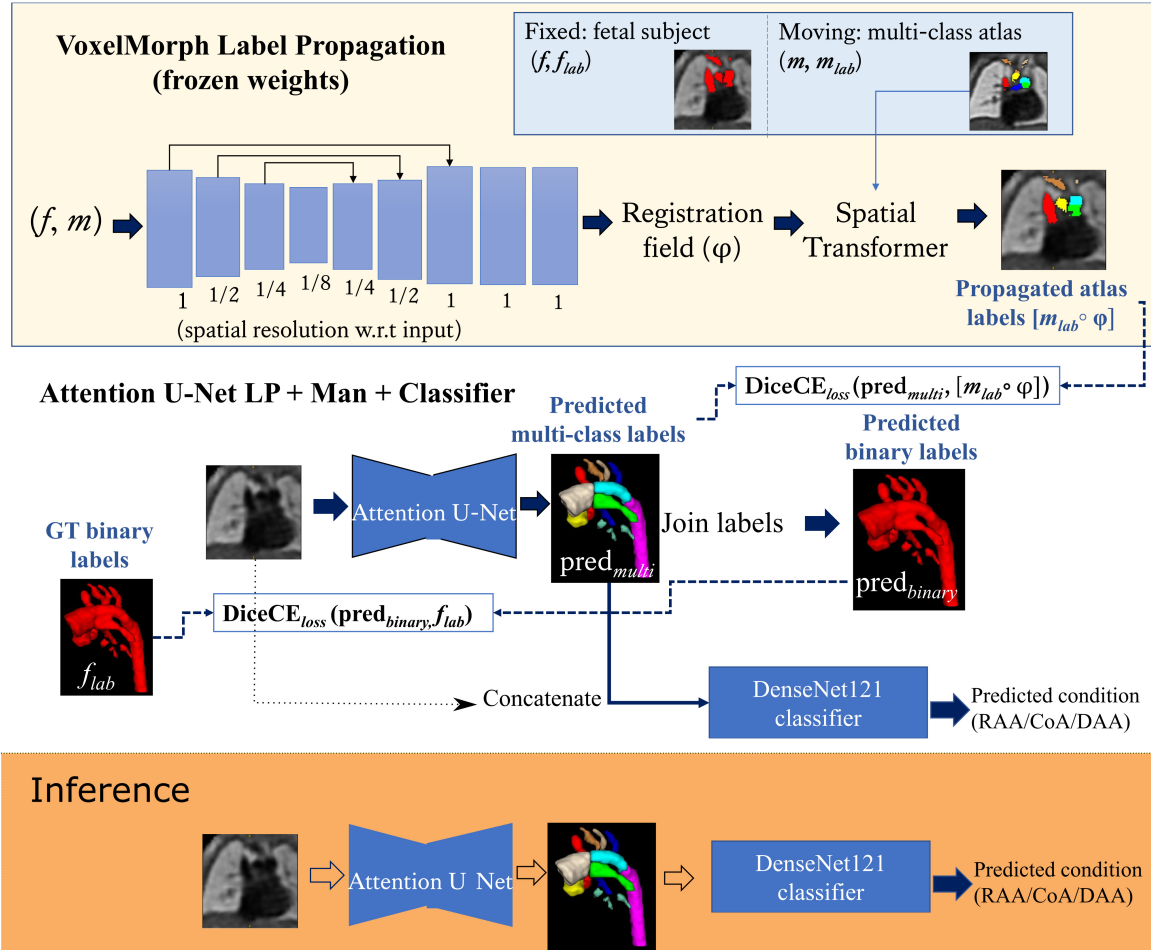2. https://github.com/BioMedIA/MIRTK

Training

Figure 4: Our proposed segmentation framework (*Attention U-Net LP + Man + Classifier*) is trained with both a multi-class propagated labels loss, and a binary loss employing the manual binary masks. We use three atlases: CoA, RAA and DAA, and train our segmentation network jointly with an anomaly classifier. No labelling information (segmentation or diagnosis) is required during inference.

2.3.1 LABEL PROPAGATION LOSS FUNCTIONS

Our similarity reconstruction loss function ($\mathcal{L}_{sim}$) is Local Normalised Cross Correlation loss (LNCC$_{loss}$) (Balakrishnan et al., 2019). We employ bending energy **BE loss** ($\mathcal{L}_{BE}$), as described in Grigorescu et al. (2020) as a displacement field regulariser. The **total registration loss** $\mathcal{L}_{reg}$ may be expressed as

$$\mathcal{L}_{reg} = \mathcal{L}_{sim}(f, m \cdot \phi) + \lambda_1 \mathcal{L}_{BE}(\phi), \tag{1}$$

where λ_1 is a loss weight.

2.3.2 Registration network implementation details

We employ a U-Net-based encoder-decoder architecture. We use blocks of 3D strided convolutions with leaky ReLU activations. We illustrate our architecture in Fig. 5.
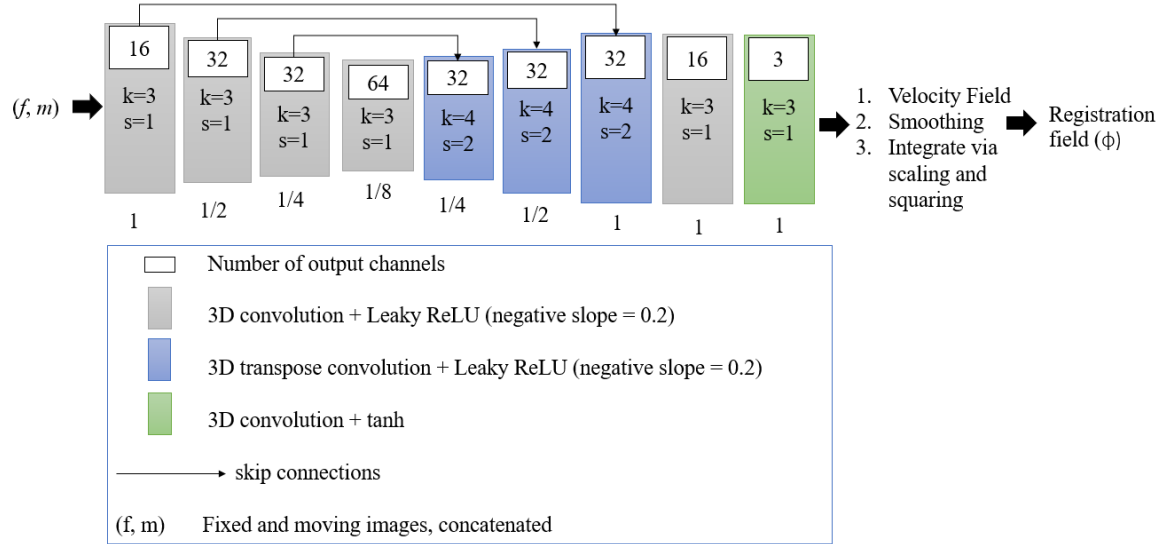

Figure 5: CNN architecture used for registration. The numbers under each convolution representation indicate the volume spatial resolution relative to the input volume. k=kernel size, s=stride.

We train a single CNN on all diagnoses, appropriately pairing each subject to its corresponding atlas. We train the VoxelMorph registration network until convergence (81,653 iterations, NVIDIA GeForce RTX 3090 GPU), using a linearly decaying learning rate initialised at 5×10^{-4} and an Adam optimiser (default β parameters, weight decay of 1×10^{-5}). We use Project MONAI spatial and intensity data augmentation[3].

2.4 CNN segmentation

3D Attention U-Net is our backbone segmentation architecture, due to its success in segmenting multi-class structures of varying locations and sizes (Oktay et al., 2018).

2.4.1 Segmentation loss function

We use the **soft dice and cross-entropy loss** (DiceCE$_{loss}$, Hatamizadeh et al. (2022)) for all our segmentation experiments, with both cross entropy and dice loss weighted equally.

We additionally investigated network performance using both **Generalised Dice loss** (GDL, Sudre et al. (2017)), and **GDL+Focal loss** (Lin et al., 2017). However, we found very poor performance when using both these losses, such as unlearnt small vessels.

Our proposed framework *Attention U-Net LP + Man* is trained using a combined loss

---

3. https://github.com/Project-MONAI/MONAI/.

$$\mathcal{L}_{seg} = \lambda_3 \text{DiceCE}_{loss}(\text{pred}_{multi}, [m_{\text{lab}} \cdot \phi]) + \\ \lambda_2 \text{DiceCE}_{loss}(\text{pred}_{joined}, f_{lab}) \tag{2}$$

where pred_{multi} are the multi-class Attention U-Net predictions, pred_{joined} are binary label predictions (multi-class output labels joined together), $[m_{\text{lab}} \cdot \phi]$ are the propagated atlas labels, f_{lab} are the manual binary labels, λ_2 is the binary loss weight, and λ_3 the multi-class loss weight. The proposed losses are schematically presented in Fig. 4.

2.4.2 Segmentation network implementation details

We use a 3D Attention U-Net (Oktay et al., 2018) implemented in Project MONAI[3] for automated segmentation, with five encoder-decoder blocks (output channels 32, 64, 128, 256 and 512), convolution and upsampling kernel size of 3, ReLU activation, dropout ratio of 0.5, batch normalisation, and a batch size of 12. We employ an AdamW optimiser with a linearly decaying learning rate, initialised at 1×10^{-3}, default β parameters and weight decay=1×10^{-5}. We use intensity and spatial augmentations from Project MONAI[3], including random intensity shifts, Gaussian smoothing, Gaussian noise, sharpening, contrast adjustments and bias field.

We train our proposed method (*Attention U-Net LP + Man*, NVIDIA GeForce RTX 3090 GPU) by increasing the binary weight (λ_2) by 0.05 every 50 epochs until convergence (15809 iterations), with the propagated labels weight set to a fixed value (λ_3=1). This is to ensure accurate vessel classification while increasing the accuracy of the whole vessels region of interest (ROI).

2.5 Classification

We train an anomaly classifier to discern between three classes: CoA, RAA and DAA. We employ DenseNet121 (Huang et al., 2017) as our backbone classifier architecture.

DenseMulti is our classifier trained jointly with a segmenter network (*Attention U-Net LP + Man*). This is a multi-task framework, where we update the weights of both networks simultaneously and use the features learnt from the segmenter to train the classifier. The inputs to the classifier are the output softmax features from the segmenter network. Fig. 6 illustrates our joint training framework. We aim to explore whether combined training will encourage the segmenter to learn the characteristic features of each anomaly.

We additionally explored training a classifier on the latent space of the segmenter network. However, we did not find a significant segmentation improvement.

2.5.1 Classification loss function

We employ different optimizers for classifier and segmenter networks and perform back-propagation using the same total loss for both networks. This is expressed as

$$\mathcal{L}_{joint} = \mathcal{L}_{seg} + \lambda_4 \mathcal{L}_{class}. \tag{3}$$

$\mathcal{L}_{seg}$ is the weighted sum of DiceCE_{loss} between propagated labels and predictions (multi-class output) and between manual binary labels and predictions (binary output), see Sec. 2.4.1. $\mathcal{L}_{class}$ is the classification cross entropy loss; and λ_4 is a tunable loss weight.

CNN segmentation

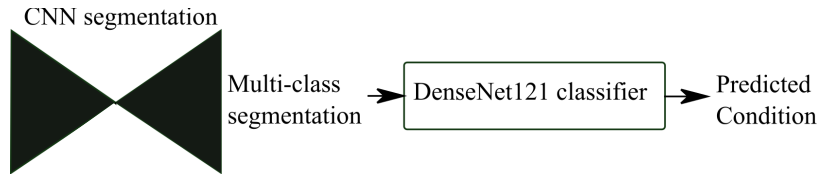Multi-class segmentation → DenseNet121 classifier → Predicted Condition

Figure 6: **DenseMulti**: our classifier experiment trained on the output softmax segmentation prediction.

2.5.2 CLASSIFICATION IMPLEMENTATION DETAILS

We pre-train our segmentation networks prior to the addition of the classifier (with both propagated and binary labels), as described in Sec. 2.4.2. The classifier is also pre-trained (on the softmax segmentation output) keeping the segmenter network weights fixed. This is to ensure that the classifier only learns highly representative features of each anomaly; and that the classifier does not degrade the segmenter network performance.

We repeat each joint training experiment three times, employing an independently trained segmentation network each time (the same repeated *Attention U-Net LP+Man* experiments described in Sec. 3.1). We use the latest binary weight (λ_2) from the best pre-trained network (on our validation set) in each case; and tune the classifier weight (λ_4 in Eq. 3), with the optimal weight being between 6 and 12 for **DenseMulti**, and 0.1 to 1 for **DenseBin** and **DenseImgMulti**, which is a classifier trained only on binary segmentation output (see Sec. 3.3). We find the weight balancing to be very important and tune this parameter for each training experiment (which employs an independently trained segmentation network each time) by observing segmentation loss convergence, classifier validation set scores and visual segmentation inspection on our validation set. We train until reaching the best performance on our validation set (lowest multi-class DiceCE$_{loss}$ on validation set).

3. Experiments

3.1 Training label type ablation study

As presented in Ramirez Gilliland et al. (2022) we conduct ablation studies on the inclusion of each type of labelling information, e.g. manual binary labels in individual subject images and multi-class labels propagated from the anomaly-specific atlases. We evaluate segmentation performance in the following experiments:

- **LP**: Label Propagation using VoxelMorph.

- **Attention U-Net LP**: Attention U-Net trained exclusively with propagated labels ($\lambda_2 = 0$ in Eq. 2).

- **Attention U-Net Man**: Attention U-Net trained exclusively with manually segmented binary labels ($\lambda_3 = 0$).

- **Attention U-Net LP+Man**: Attention U-Net trained with both manual binary labels and propagated labels (Eq. 2).

Expanding on our previous work (Ramirez Gilliland et al., 2022), we repeat our segmentation experiments three times, to account for network stochasticity.

3.2 Weak supervision

We study the impact a varying amount of manually generated labels have on segmentation network performance. In this ablation study, we solely examine the weak binary manual labels as they are meticulously created by expert fetal cardiac clinicians, while the propagated labels are automatically derived.

Six segmentation experiments are compared, with varying percentages of manual binary labels (Man) included: 0% Man (no manual binary labels in training, i.e. *Attention U-Net LP*), 5% Man, 25% Man, 50% Man, 75% Man, and 100% Man (i.e. our proposed *Attention U-Net LP + Man*). These percentages are computed for each anomaly (out of their respective total number of training cases). In all these experiments the propagated labels are fully used. The same implementation details as described in Sec. 2.4.2 are employed.

3.3 Classification experiments

We conduct four main experiments to evaluate our aortic arch anomaly classifiers.

DenseMulti is a DenseNet121 trained from the softmax output of the **multi-class** segmenter network (*Attention U-Net LP + Man*, see Fig. 6).

DenseBin is a DenseNet121 trained from the softmax output of our **binary** segmenter network (*Attention U-Net Man*).

DenseImage, is a DenseNet121 trained exclusively on the volume images as input.

DenseImgMulti, is a DenseNet121 trained on the multi-class softmax output of *Attention U-Net LP + Man*, concatenated with the input T2w image.

We pre-train **DenseBin**, **DenseMulti** and **DenseImgMulti** using the frozen weights of each respective segmenter network (*Attention U-Net Man*, and *Attention U-Net LP + Man*). We report these results under each classifier name followed by **Separate**. We then train our pre-trained classifier and segmenter networks jointly, which is our multi-task approach, and report results under the title **Joint**.

We repeat each experiment three times, employing our three independently pre-trained segmenter network rounds.

3.4 Evaluation metrics

3.4.1 Quantitative analysis

We manually generated multi-class ground truth (GT) labels for our test set (N=40) via ITK-SNAP (Yushkevich et al., 2006). We report similarity metrics including multi-class Dice scores, recall, precision, average surface distance (ASD), and 95th percentile of the Hausdorff Distance (HD95). We repeat our segmentation experiments three times and average results. We also include a short analysis studying the impact of image quality on performance.

3.4.2 Qualitative analysis

There are inherent limitations to conducting a quantitative analysis exclusively, particularly given the quality of our data and the task at hand. We find the topological correctness of the anomaly area, our key segmentation objective, to not be reflected in our quantitative metrics described above. We thereby devise a **qualitative analysis** strategy, which involves manual inspection and scoring of the anomaly area of each test set subject.

We assess the topological correctness of the aortic arch (anomaly area), as described in Table 2, where a score of 1 (best) represents an aortic arch with topologically correct anomaly delineation, and a score of 3 (worst) indicates topologically incorrect delineation. Types of topological errors for segmentations with a score of 3 include split aortic arch, indiscernible aortic arch (merged), an unsegmented or partially segmented right arch in a DAA case (split arch), a segmented left arch for RAA cases (double arch segmented), and a segmented right arch for CoA cases (double arch segmented). The term oversegmentation refers to erroneously labelling background voxels as an anatomical structure, while rarely labelling structure as background. This typically results in erroneously thick vessels.

Table 2: Description of our anomaly area topological assessment. HN refers to head and neck vessels, and AD arterial duct.

	Aortic arch score
1	Topologically correct aortic arch
2	Oversegmented aortic arch merging into HN vessels or AD.
3	Topologically incorrect aortic arch

4. Results

4.1 Training label type ablation study

Here we present metrics comparing our proposed segmentation framework which learns from both binary individual labels and multi-label atlases, against the inclusion of just one type of labelling (see Sec. 3.1). Differences to Ramirez Gilliland et al. (2022) are due to adaptations made to the atlas in pulmonary arteries and exclusion of pulmonary veins (following discussions with clinicians), and stochasticity due to network retraining. All metrics displayed are on unseen test set, comparing to our manually generated GT.

4.1.1 Whole vessel ROI

Here we present metrics for the whole vessels' ROI in violin plots in Fig. 7. In the case of multi-label segmentation, the vessels were joined into a single region prior to the metrics calculations.

Attention U-Net Man, which is the network trained on binary labels, displays the highest mean vessels' ROI dice (0.83 ± 0.03). However, further analysis reveals a much higher average recall (0.87 ± 0.05) than precision (0.79 ± 0.07), which points to significant oversegmentation. This is clearly visible in Fig. 8. We can observe that oversegmentation results in the merging of the neighbouring vessels, obstructing the visualisation of the fetal cardiac vessels' anatomy.

Conversely, the proposed network *Attention U-Net LP + Man* has a more balanced performance with a mean Dice $= 0.79 \pm 0.03$, a mean precision $= 0.80 \pm 0.05$, and a mean recall $= 0.78 \pm 0.05$. The resulting segmentations (Fig. 8) offer a clear depiction of individual vessels, which is required for our clinical application.

Attention U-Net Man also has the lowest mean ASD of 0.51 ± 0.091 mm, which increases to 0.55 ± 0.08 for the proposed *Attention U-Net LP + Man*. However, *Attention U-Net Man* presents outliers with ASD even greater than *LP*, whereas *Attention U-Net LP + Man* is much more consistent across cases.

The network trained on propagated labels only *Attention U-Net LP* has the lowest performance in all the metrics.
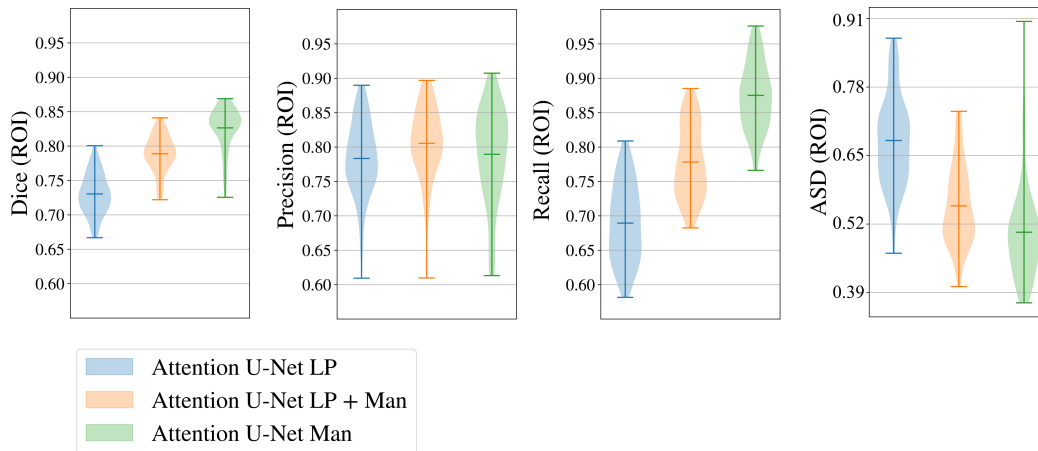
Figure 7: Quantitative ROI metrics for our labelling ablation experiments. The horizontal lines represent the means.

4.1.2 Individual fetal cardiac vessels

Table 3 displays metrics comparing our multi-class experiments. We compare our full approach (*Attention U-Net LP + Man*) to Attention U-Net trained exclusively on propagated labels (*Attention U-Net LP*) and VoxelMorph label propagation (*LP*). These results indicate that adding the binary labels to the training of the segmentation network offers a significant improvement for Dice and ASD metrics for all vessels. Further, Attention U-Net trained on propagated labels improves on label propagation alone.

4.1.3 Visual inspection

Nevertheless, we find these quantitative metrics to not be fully descriptive of the anatomical correctness of the segmentation, a decisive feature when segmenting anomalies. Following visual inspection, we can conclude two main advantages of our multi-class approach: **At-tention U-Net LP + Man is more consistent in both small vessel detection and topological correctness of the anomaly area compared to *Attention U-Net Man*.** The latter is quantified in Sec. 4.4. Fig. 8 includes a depiction of these two prominent issues

Table 3: Similarity metrics between our multi-class experiment predictions and our manually generated GT. We highlight the best scores in bold. HN indicates the head and neck vessels (LSA, LCCA, BCA, RSA). The three bottom rows report metrics averaged over all vessels. We highlight with an * the vessels with a p-value< 0.05 (Mann-Whitney U test) (non-parametric) compared to *Attention U-Net LP + Man.*

	LP		Attention U-Net LP		Attention U-Net LP + Man	
Vessel	Dice	ASD	Dice	ASD	Dice	ASD
SVC	0.65 (0.15)*	0.88 (0.60)*	0.70 (0.06)*	0.70 (0.20)*	**0.76 (0.06)**	**0.56 (0.16)**
LPA	0.58 (0.15)*	0.88 (0.93)*	0.63 (0.08)*	0.72 (0.30)*	**0.65 (0.08)**	**0.68 (0.29)**
RPA	0.55 (0.12)*	0.80 (0.44)*	0.61 (0.07)*	0.71 (0.23)	**0.63 (0.07)**	**0.67 (0.21)**
Aortic arch	0.58 (0.13)*	0.87 (0.46)*	0.62 (0.06)*	0.79 (0.16)*	**0.76 (0.04)**	**0.54 (0.08)**
AD	0.70 (0.16)*	0.66 (0.57)*	0.76 (0.07)*	0.49 (0.11)*	**0.78 (0.07)**	**0.44 (0.12)**
DAO	0.77 (0.11)*	0.63 (0.40)*	0.81 (0.04)*	0.52 (0.20)*	**0.84 (0.04)**	**0.43 (0.18)**
MPA	0.74 (0.11)*	0.71 (0.25)*	0.76 (0.06)*	0.68 (0.16)*	**0.81 (0.05)**	**0.55 (0.13)**
HN	0.31 (0.21)*	1.50 (1.27)*	0.38 (0.20)	1.24 (0.94)*	**0.40 (0.19)**	**1.20 (0.93)**
Avg. Dice	0.54 (0.24)		0.59 (0.20)		**0.63 (0.21)**	
Avg. HD95	3.14 (1.07)		2.71 (1.81)		**2.50 (1.87)**	
Avg. ASD	2.26 (0.97)		0.85 (0.63)		**0.77 (0.63)**	

observed across *Attention U-Net Man* experiments: undetected small vessels, and merging of the aorta into surrounding vessels.
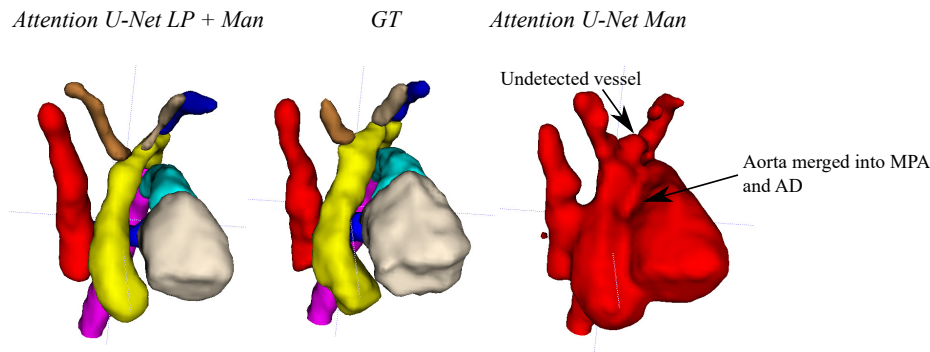
Figure 8: Depiction of aorta topological inaccuracies and undetected small vessels commonly occurring in *Attention U-Net Man.*

4.1.4 VISUALISATION OF THE LATENT REPRESENTATION

We inspect the anomaly-specific discernment of our labelling ablation networks by visualising t-SNE reduced latent space features (Fig. 9). With this, we can easily examine the effect multi-class and binary labels have on anomaly distinction for our networks.

We can conclude that our multi-class network *Attention U-Net LP + Man* offers highly clustered groups for anomalies even in the bottleneck features, contrarily to our binary network (*Attention U-Net Man*) which especially struggles to discern between RAA and
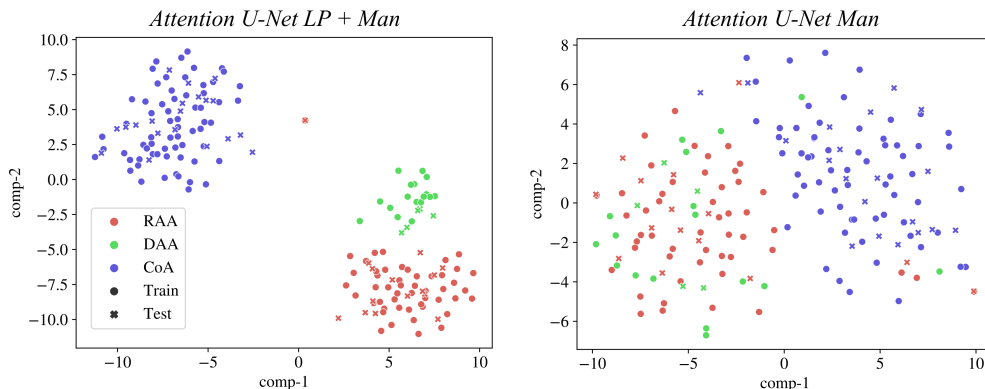
Figure 9: t-SNE bottleneck reduced features for one training round of our combined label framework (multi-class, *Attention U-Net LP + Man*) and our binary network (*Attention U-Net Man*).

DAA cases, as supported by our qualitative analysis (Sec. 4.4). We include further t-SNE plots comparing our various experiments in Appendix B.2.

4.1.5 WEAK SUPERVISION ABLATION

Fig. 10 includes Dice and ASD scores for the vessels' ROI and aortic arch label, for an increasing number of training cases with manually generated binary labels. Our findings indicate that including binary labels in training results in a statistically significant improvement in segmentation performance (p-value < 0.05, Mann-Whitney U test) compared to not including any binary labels (0%, *Attention U-Net LP*) when 25% or more of the manual labels were included. For 5% of included manual labels, the improvement in Dice and ASD is only significant for the aortic arch region, not the whole ROI.

4.2 Classification of anomalies

4.2.1 CLASSIFICATION PERFORMANCE

Table 4 presents accuracy and balanced accuracy scores (average recall obtained on each class) for each classifier experiment, as well as recall scores for each individual anomaly. In addition to comparing different types of features for classification (multi-label segmentation, binary segmentation and original image), we also compare separate and joint training of segmenter and classifier networks. The confusion matrices are provided in Appendix B.

We observe that while joint training of the classifier with binary segmentation improves the classification results, this is not the case for multi-class segmentation. For multi-class segmentation, the differences in classification performance in jointly and separately trained frameworks are caused by a small number of subjects, with three misclassifications (two individual subjects) over the three training rounds in the joint framework (one misclassification per round), and two misclassifications in a single subject in the separate framework. There is therefore no clear improvement with joint training in the multi-label case, however,

(a) ASD scores (lower = better).
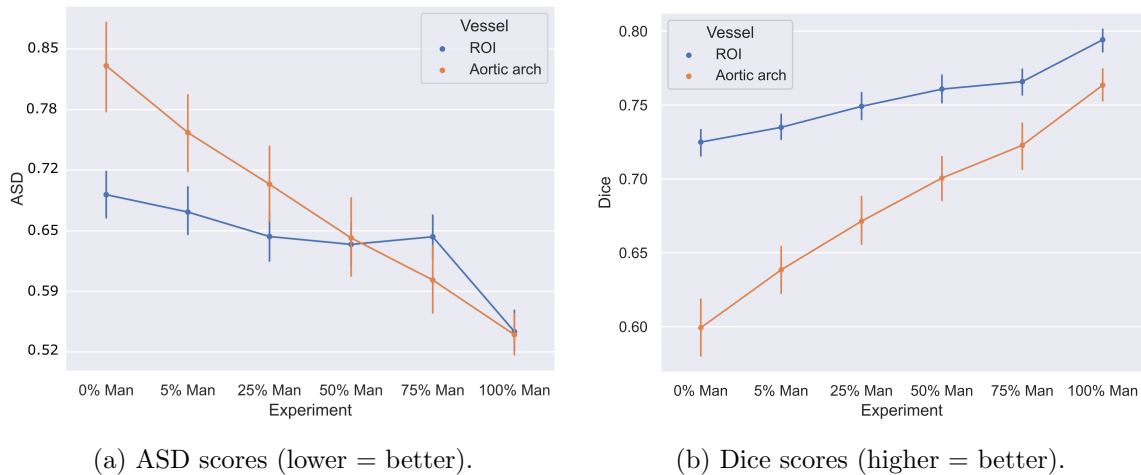
(b) Dice scores (higher = better).

Figure 10: Test set metrics for the vessels' ROI, and aortic arch predicted label comparing the impact of adding binary manual labels (Man) during training. Error bars represent the 95% confidence interval, computed via bootstrapping (non-parametric uncertainty representation).

the classification performance is high, with an accuracy of 0.97 for **DenseMulti** and 0.99 for **DenseImgMulti**.

In general, CoA cases are always correctly classified, while the misclassifications happen between RAA and DAA cases.

Overall, the anomaly classification from multi-label segmentation clearly outperforms the classification from binary segmentations or directly from the images. However, when incorporating image information concatenated with the output segmentation, the classifier achieves the highest performance. This image data inclusion allows the network to focus on global features and becomes less reliant on topologically accurate segmentation predictions compared to **DenseMulti**. Consequently, **DenseImgMulti**'s high accuracy is not truly reflective of segmentation performance, as correct classification can still occur even with important topological errors in the anomaly area.

We display our **DenseMulti** class probabilities in Fig. 11 for each anomaly (one vs. rest). These include the results of our three training rounds. We observe a higher number of cases with a lower softmax probability for RAA and DAA groups, meaning that our classifier is less confident when distinguishing between RAA and DAA and the rest of our groups, however very confident for CoA predictions. This is understandable, given the anatomical similarities between RAA and DAA subjects (Fig. 2).

4.2.2 Analysis of misclassified cases

Here we provide further detail on our **DenseMulti** and **DenseImgMulti** misclassifications. We obtain one misclassified case per round with **DenseMulti**, with one case repeated across two of our training rounds. This same case is the one misclassification for **DenseImgMulti**. This case is an RAA subject erroneously classified as DAA due to the double arch being wrongly segmented (Fig. 12). This particular case presented lower visibility in the anomaly

Table 4: Accuracy and recall scores for our classifier experiments, averaged over all three training rounds. Separate = prior to joint segmenter and classifier training, Joint = after joint training. Bl. Acc. = balanced accuracy.

	Accuracy	Bl. Acc.	Recall CoA	Recall RAA	Recall DAA
DenseMulti					
Separate	0.98 (0.01)	0.95 (0.04)	**1.0 (0.0)**	**1.0 (0.0)**	0.87 (0.11)
Joint	0.97 (0.0)	0.98 (0.0)	**1.0 (0.0)**	0.93 (0.0)	**1.0 (0.0)**
DenseBin					
Separate	0.88 (0.01)	0.90 (0.01)	0.95 (0.0)	0.75 (0.04)	**1.0 (0.0)**
Joint	0.92 (0.02)	0.91 (0.05)	0.97 (0.03)	0.89 (0.10)	0.81 (0.17)
DenseImgMulti					
Separate	**1.0 (0.0)**	**1.0 (0.0)**	**1.0 (0.0)**	**1.0 (0.0)**	**1.0 (0.0)**
Joint	0.99 (0.01)	0.99 (0.01)	**1 (0.0)**	0.98 (0.04)	**1.0 (0.0)**
DenseImage	0.82 (0.07)	0.79 (0.09)	0.98 (0.03)	0.64 (0.01)	0.62 (0.30)
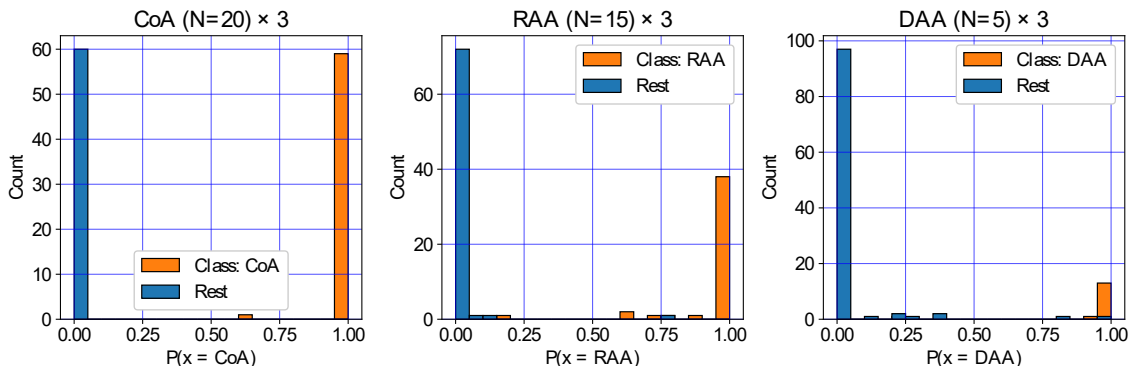
Figure 11: Count of probability distributions for all three rounds of **DenseMulti** classifier experiment on our test set.

area, with more than one independently trained network partially segmenting (erroneously) the left arch. We highlight the blurriness in this region causing the oversegmentation. This particular case already presented topological issues prior to adding the classifier (Fig. 12, left). The addition of the classifier clearly tries to erroneously correct this initial error, by fully segmenting the double arch, i.e. by pushing the segmentation prediction closer to one of the aortic arch anomalies. Importantly, this case presents topological issues for one of our joint **DenseImgMulti** experiments but was correctly classified. These findings suggest **DenseImgMulti** to be a more reliable classifier when the segmentation is faulty, as it incorporates image information. Consequently, it cannot be utilized as a quality control tool to identify erroneous segmentations.

Our second misclassification is another RAA case mislabelled as DAA due to poor image quality and misalignment to the atlas (Fig. 13). Our segmentation here is accurate, however, the aortic arch and AD are highly elevated compared to the atlas (Fig. 13). This may have
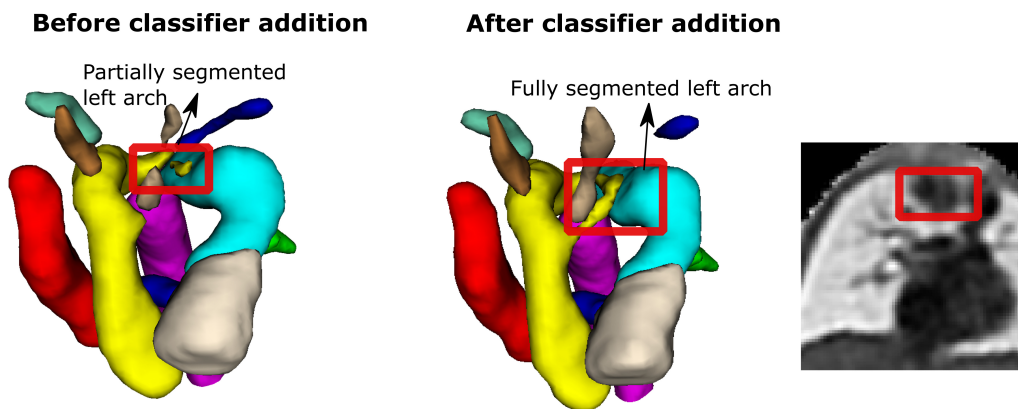
Figure 12: Arrow signals the double arch erroneously segmented in an RAA subject, resulting in misclassification. We highlight the blurrier region where this error occurs in the image with a rectangle.

been misclassified after joint training due to slight overfitting to the training data. However, it is only one case (out of 120 in total).
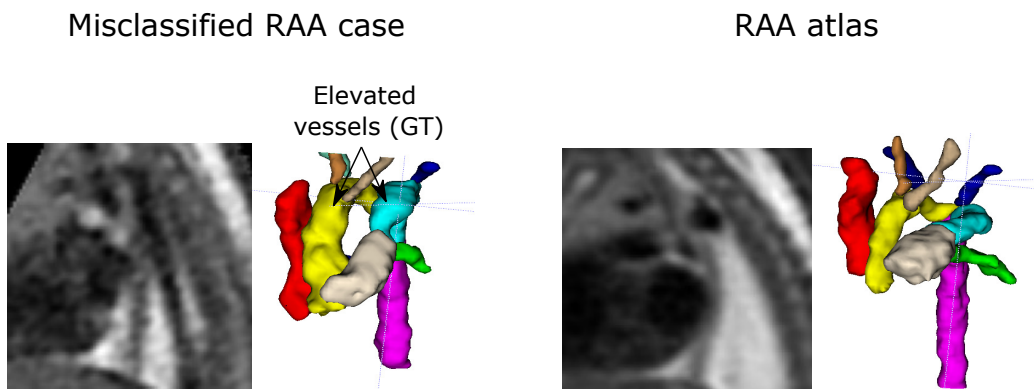

Figure 13: Mislabelled RAA as DAA in one of our rounds due to low image quality and misalignments to the atlas, with a highly elevated aorta and AD compared to the atlas.

4.2.3 SEGMENTATION IMPROVEMENTS DUE TO ANOMALY CLASSIFIER

Regarding segmentation network gains, we find meaningful improvements in our multi-class framework, particularly regarding the completeness of double arch segmentation (Figs. 15 and 14). Although our DAA test set comprises only five cases, we repeat this experiment three times, with independently trained segmenter networks, and find that in **all correctly classified DAA subjects where the right arch was originally partially segmented or unsegmented, adding a classifier (DenseMulti and DenseImgMulti)** resolved this. The double arch vessel thickness is also improved. This constitutes **eight subjects with a notable improvement over the three rounds**. We report topological improvements in our overall qualitative analysis, Sec. 4.4, Fig. 19.

Fig. 14 depicts a clear example of improved segmentation performance via our multi-task framework, consequently correcting the predicted anomaly class.
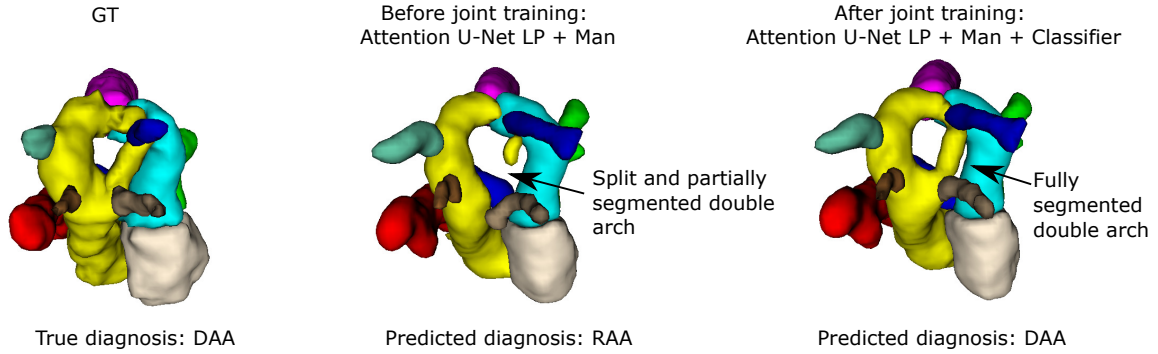

Figure 14: Test case where we find a significant topological improvement of the double arch, resulting in a corrected anomaly prediction.
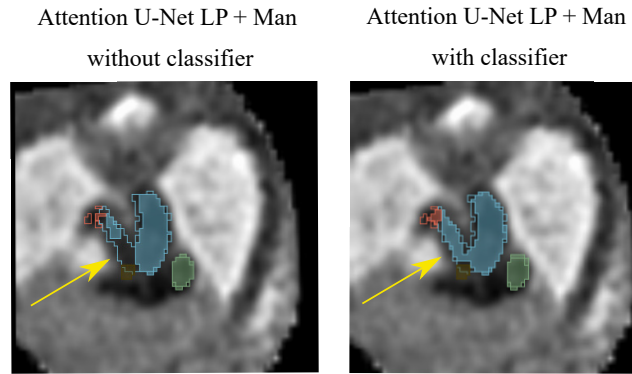

Figure 15: Arrow signals the double arch in a DAA subject, a large portion unsegmented prior addition of the classifier network. The GT is overlaid without filling, i.e. just the contours, while the predictions are filled.

We find marginal improvements in aorta metrics for DAA cases (see Tab. 5). The fact that these improvements are marginal is likely due to the small surface area the anomaly location comprises, despite being key for robust segmentation.

Table 5: Marginal improvements in Dice and ASD metrics of the aorta, observed after our joint training. These contain our averaged results for all three training rounds. We highlight in bold our best metrics (higher=better for dice, lower=better for ASD).

| | Attention U-Net LP + Man | | Attention U-Net LP + Man + Classifier | |
	Aorta Dice	Aorta ASD	Aorta Dice	Aorta ASD
CoA	0.74 (0.04)	0.82 (0.15)	0.74 (0.04)	**0.81 (0.14)**
RAA	0.77 (0.02)	**0.83 (0.10)**	0.77 (0.03)	0.85 (0.11)
DAA	0.76 (0.02)	0.91 (0.09)	**0.77 (0.02)**	**0.87 (0.07)**

Regarding our binary framework (*Attention U-Net Man*), we find that adding a classifier (**DenseBin**) affects the segmentation output with both improvements observed in double arch segmentation (Fig. 16) and degradation in certain cases such as heightened vessel merging.
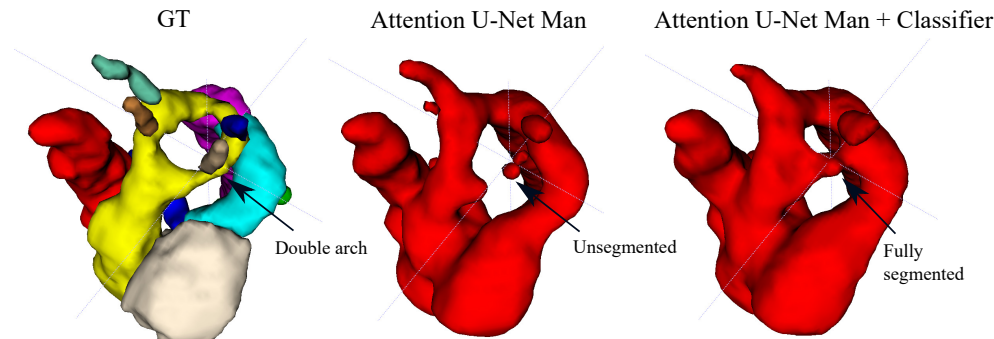

GT   Attention U-Net Man   Attention U-Net Man + Classifier

Figure 16: Improvements observed in double arch segmentation after adding a classifier (**DenseBin**) to our binary framework.

We postulate that this difference in performance is due to both (a) the anomaly classifier being presented with a more challenging task (harder to identify aorta and anomaly area from a single vessel), and (b) labelling variations due to inter-observer variability, contrasting with our propagated atlas labels.

4.3 Image quality analysis

In order to study the impact image quality has on our predictions, a concise analysis of image quality against quantitative network performance is included here.

Two factors are considered: the reconstruction process (SVR Kuklisova-Murgasova et al. (2012); Kainz et al. (2015) or DSVR Uus et al. (2020)), and a manual image quality assessment. The latter consists of manually revising and scoring the test set images from 1-3 (best to worst), based on image noise, visibility, and vessel sharpness, with 1 being the highest quality, 2 being average quality, and 3 being poor quality and visibility.

A note of consideration is that our test set is majoritarily formed of SVR reconstructions (lower quality, N=102 across all three repeated training rounds), with only 18 DSVR cases.

Fig. 17b includes precision and sensitivity scores of the vessels' ROI comparing performance for our manual image quality assessment, with Fig. 17a depicting Dice scores for the vessels' ROI and aortic arch region. We generally observe a higher number of lower-scoring outliers in lower-quality images (score = 3), however, we find differences between groups to be statistically non-significant.

Regarding performance on image reconstruction type (Fig. 18), we find a statistically significant difference (p-value = 3.97×10^{-4} for a Mann-Whitney U test) in performance in HD95 of the vessels' ROI (Fig. 18b), with DSVR cases presenting lower scores (better performance).

Concerning classifier performance, we find all misclassifications from **DenseMulti** and **DenseImgMulti** to be exclusively SVR reconstructions.
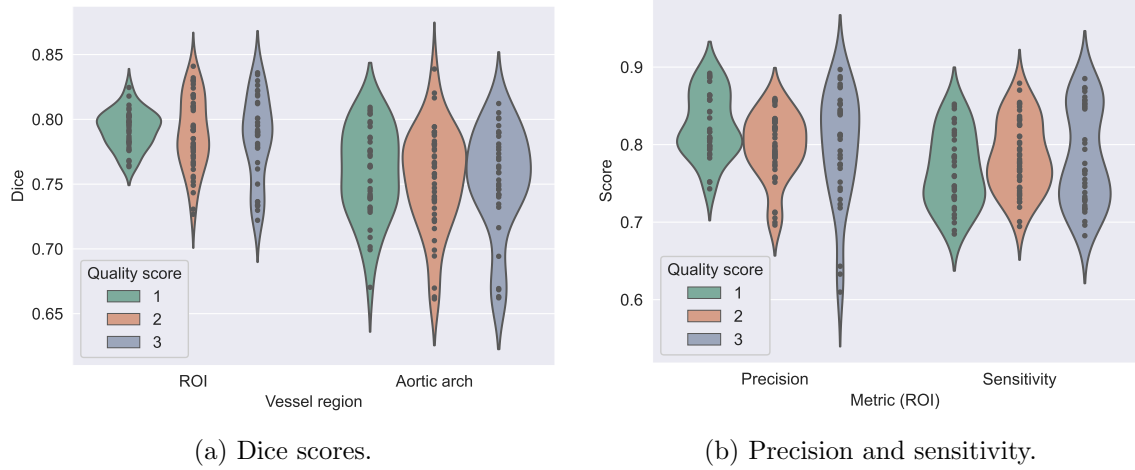
(a) Dice scores.

(b) Precision and sensitivity.

Figure 17: Test set metrics comparing the impact image quality ($1 \rightarrow 3 = \text{best} \rightarrow \text{worst}$) has on network performance.

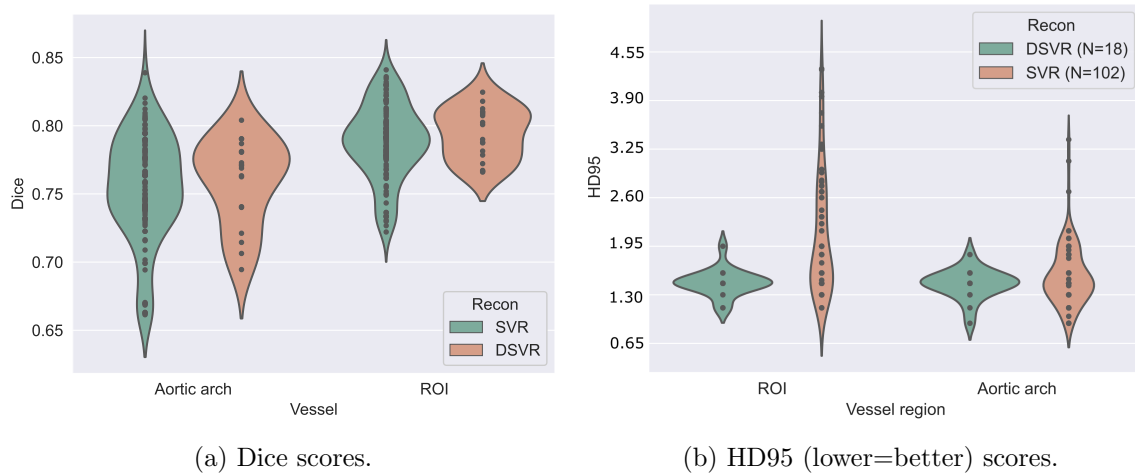(a) Dice scores.

(b) HD95 (lower=better) scores.

Figure 18: Test set metrics comparing performance for DSVR and SVR reconstructions.

4.4 Qualitative results

We find that the quantitative metrics presented in previous sections do not always capture the topological and anatomical correctness of the anomaly area, which is key for our clinical application. Accordingly, we present a qualitative topological anomaly area analysis for our segmentation predictions (Sec. 3.4 Tab. 2). Fig. 19 compares our main ablation studies for classifier inclusion and labelling information for all diagnoses, with our full framework being **Attention U-Net LP + Man + Classifier**. We include the performance on DAA test set exclusively in Fig. 19 left. See Appendix B.1 for similar CoA and RAA plots.

Figs. 19 clearly demonstrate the added value of using multi-class propagated labels, as *Attention U-Net LP + Man* increases the number of topologically correct segmentations over *Attention U-Net Man* by 68%.
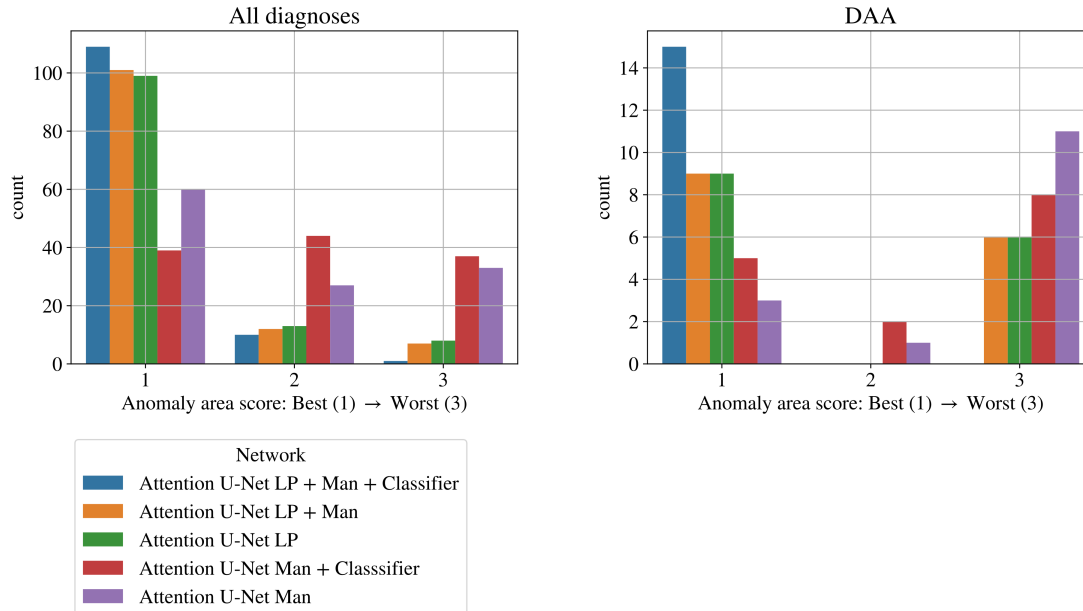
428

Figure 19: Qualitative analysis results on topology correctness of the anomaly area (aortic arch) on our test set for **all diagnoses** (left), and for DAA cases only (right). Each training experiment was repeated three times, and we include all of our results here (i.e. we have three sets of results for each of our 40 test set cases, one for each round, so the total number of values for each technique is 120)

.

The top performing experiment regarding the topological correctness of the anomaly area is _Attention U-Net LP + Man + Classifier_ (DenseMulti), our final framework. We find important improvements by adding a classifier to our multi-class framework, particularly regarding cases with a score of 3. These improvements are predominantly DAA cases which initially present a split or unsegmented double arch, corrected by adding a classifier (see Fig. 19 (right) and Sec. 4.2.3).

Overall, we still observe some cases with a score of 2, and one with a score of 3 (which includes misclassified subjects) in our full framework (_Attention U-Net LP + Man + Classifier_). We inspect these cases and find that they are the same three subjects with poor segmentation performance across the three experiment training rounds, presenting both poor image quality and misalignments to the atlases. See Appendix B.1 for extended analysis and visualisation of these cases.

An important disadvantage of the binary segmentation network is vessel merging, in extreme cases producing an indiscernible aorta (Fig. 20). Interestingly, we find the addition of a classifier (**DenseBin**) to our binary framework (_Attention U-Net Man + Classifier_) to generally degrade performance by heightening this aortic merging problem. Nonetheless, we do observe improvement in certain correctly classed DAA cases (see Sec. 4.2.3), albeit not being as consistent and robust as our multi-class improvement.
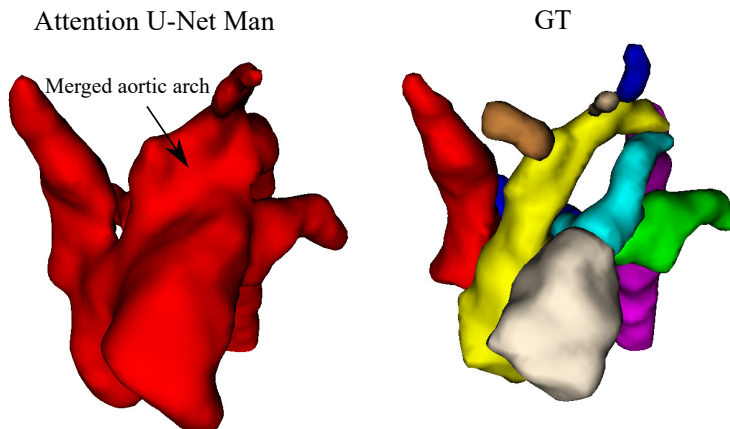
Figure 20: Extreme test set case with aortic merging issue observed.

5. Discussion

We present a multi-task deep learning framework for multi-label fetal cardiac vessel segmentation and anomaly classification in T2w 3D fetal MRI.

Fetal CMR has only recently demonstrated its potential (Lloyd et al., 2019), due to the highly specific challenges this data presents. Given this, the field standards regarding automated segmentation performance have yet to be set. Notably, our dataset requires fast acquisition protocols and motion correction and reconstruction algorithms (Uus et al., 2020), due to the inherent motion corruption present in fetal imaging (fetal motion, rapid fetal heartbeat, maternal motion), as well as poor resolution and contrast.

Additionally, there is a range of anatomical variability within each anomaly, as well as differing vessel sizes, from the descending aorta which is easily discernible to small head and neck vessels which are sometimes absent even in the GT due to low visibility.

Our specific dataset presents a range of image qualities. We manually score our test set images (based on vessel visibility, artefacts and noise), to provide further explainability into potential failed cases. Our test set comprises 12 high quality subjects, 16 average quality, and 12 low quality images. Therefore, a high segmentation scoring performance is extremely challenging, particularly for small vessels in low quality images, or for cases with important anatomical variability.

5.1 Segmentation training label type ablation

Our labelling experiments highlight two important points: (1) adding manual labels substantially increases network performance, and (2) multi-class information is crucial for small vessel detection, as well as anomaly area segmentation. We find significantly increased similarity metrics after including the binary manual labels, with a similar qualitative topology score.

We find important anatomical inconsistencies in our binary segmentations, including merged vessels, contrary to our proposed multi-class networks. We postulate that this is due to our multi-class network learning each label individually, therefore ensuring that each vessel is present and distinguishable in the final network. The consistency in our

propagated atlas labels may also contribute towards this, which contrasts with the inter-observer variability present in the binary manual labels.

The oversegmentation we observed both qualitatively and quantitatively from our binary network is detrimental to clinical usability, as this causes vessel merging which creates indistinguishable vessel topology.

An important point from our analysis is that we find commonly used quantitative metrics to not be fully descriptive of segmentation performance regarding topological and anatomical correctness, hence our qualitative evaluation. Therefore there is a need for future work to explore more fitting topological metrics (Hu et al., 2019). The inclusion of topology-aware losses (Clough et al., 2020; Byrne et al., 2021; Clough et al., 2019; Hu et al., 2019) is also an important aspect for future work to interrogate, given the importance of generating topologically correct predictions for our task.

Noteworthy, the segmentation task in the present paper could have been tackled following alternative training strategies. A potentially simple solution could be to leverage propagated labels to train a splitting network to convert binary labels to multiple labels. Appendix A contains two additional experiments, that utilise such splitting network. In the first experiment, the splitting network is applied to binary labels predicted by *Attention U-Net Man*, while in the second experiment, the splitting network converts manual binary labels to multi-label segmentations to provide training data for a simple supervised Attention U-Net. We found these strategies to generally lead to more topological errors in vessel segmentations than for our proposed framework, see Fig. 24. In particular, we observed more mislabelled vessels in regions with poor visibility due to imaging artefacts.

5.2 Weak supervision

The weak supervision ablation study demonstrates that, although our manually generated labels are not fully labelled (as the vessels are represented as a single entity, as opposed to multi-class labels), their addition to the segmenter training framework improves segmentation performance. Even the addition of manually generated labels to only 5% of the training data (for each diagnosis) supposes a statistically significant improvement for the aortic arch label.

The highest performance is achieved via our proposed segmentation framework: *Attention U-Net LP + Man*.

5.3 Classification of anomalies

We find one explainable mislabelled test case per training round for our **DenseMulti** classifier (multi-class), and one misclassification overall for **DenseImgMulti**, with notable improvements in the anomaly area segmentation after joint network training. The addition of a classifier to our multi-class segmentation pipeline consistently improves the topological correctness of all correctly classified DAA cases with segmentation issues, as reflected by our qualitative analysis. It is important to note that the stability of our multi-task training framework depends on accurate hyperparameter tuning of the loss weight balancing in each experiment.

Prior knowledge of the anomaly will always exist in a clinical situation, and therefore misclassifications in **DenseMulti** aid the identification of cases with faulty segmentation

predictions. **DenseImgMulti**, although being the most accurate classifier, would therefore be less relevant to this application, given its inability to detect topologically incorrect segmentation predictions. It is therefore more suitable as a diagnostic support tool.

An inherent limitation of our approach is the fact that the classifier learns distinctive anomaly-specific features, which restricts the flexibility of the segmentation topology. Therefore, this framework may not generalise well to data which deviate strongly from the training data regarding image quality or anatomical variations. The generalisability of our framework to out-of-distribution cases is therefore a valuable avenue for future work to explore.

The worst performing cases in our test set comprise lower quality SVR reconstructions, that only correct rigid motion, and therefore do not account for deformations in fetal movement, which may result in blurring of the reconstructed 3D images. This was combined with noticable anatomical deviation from the atlases. We find one extreme case where the addition of a classifier accentuated an initial segmentation issue, by fully segmenting the left arch in an RAA case (Fig. 12). This is an extreme case where we observe the limitation of somewhat topologically constraining our segmentation to a set of learnt distinct features: if there is a low quality case, with poor visibility, the addition of the classifier will push the segmentation towards pertaining to one specific anomaly. Recent advancements in reconstruction techniques (Uus et al., 2020) may ameliorate this problem, as we find that all of our high quality DSVR reconstructions achieve high accuracy in both segmentation and anomaly classification.

Highly consistent multi-class labels are a key contributor towards our success. We observe a performance degradation in our binary framework (*Attention U-Net Man + Classifier*), likely due to the more challenging anomaly classifier task: the aorta is not classed as an individual vessel, and the anomaly area occupies a relatively small region in the whole vessels' ROI label. Additionally, the consistency divergence between manually segmented binary labels (inter-observer variability, shortened vessel length for lower quality images) and propagated atlas labels, as well as the lower-quality segmentation predictions of *Attention U-Net Man* (vessel merging and undetected head and neck vessels) contribute towards making our classifier (**DenseBin**) training data more varied. This impedes the classifier from learning representative aortic arch anomaly features.

Although our work is developed to suit a very specific dataset situation (DSVR and SVR fetal CMR reconstructions, anomaly-specific atlases, and partially labelled training set), our joint classification and segmentation framework may be applicable to many environments with a dual task, or for segmentation improvements in a network trained on distinctly varied data. We note also that our task has highly specific challenges, and therefore success in our application is promising for generalisation to less challenging tasks (e.g. fewer label classes, larger segmentation area, less diverging anomalies). Additionally, our work addresses a partially labelled dataset, largely common in the medical imaging field.

6. Conclusion

We present a multi-task approach for joint fetal cardiac vessel segmentation and aortic arch anomaly classification from T2w 3D MRI. We combine deep learning label propagation from anomaly-specific atlases with Attention U-Net segmentation and DenseNet121 classification.

We demonstrate a potential application of our segmentation tool for automated diagnosis, by jointly training an anomaly classifier on our output segmentations. Our multi-task approach improves the anomaly area segmentation performance, providing both multi-class segmentations and aortic arch anomaly classification. Our strategy is simple and innovative, and our work has strong clinical applicability, being highly novel regarding our dataset. Our clinical contribution is both to aid vessel visualisation for clinical reporting purposes and to aid diagnostic confidence and efficiency. Our automated multi-task tool may also be useful in training non-expert clinicians, which is challenging due to the highly specific cardiac anomalies, low dataset visibility, and anatomical variability between cases.

Acknowledgments

Ethical Standards

All fetal MRI datasets used in this work were processed subject to informed consent of the participants [REC: 07/H0707/105; REC: 14/LO/1806].

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

The individual fetal MRI datasets used for this study are not publicly available due to ethics regulations.

References

Paul Aljabar, Rolf A Heckemann, Alexander Hammers, Joseph V Hajnal, and Daniel Rueckert. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage*, 46(3):726–738, 2009.

Arghavan Arafati, Peng Hu, J Paul Finn, Carsten Rickers, Andrew L Cheng, Hamid Jafarkhani, and Arash Kheradvar. Artificial intelligence in pediatric and adult congenital cardiac mri: an unmet clinical need. *Cardiovascular diagnosis and therapy*, 9(Suppl 2): S310, 2019.

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

Kate L Brown, Deborah A Ridout, Aparna Hoskote, Lynda Verhulst, Marco Ricci, and Catherine Bull. Delayed diagnosis of congenital heart disease worsens preoperative condition and outcome of surgery in neonates. *Heart*, 92(9):1298–1302, 2006.

Nick Byrne, James R Clough, Giovanni Montana, and Andrew P King. A persistent homology-based topological loss function for multi-class cnn segmentation of cardiac mri. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*, pages 3–13. Springer, 2021.

Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in Cardiovascular Medicine*, 7:25, 2020.

James R Clough, Ilkay Oksuz, Nicholas Byrne, Julia A Schnabel, and Andrew P King. Explicit topological priors for deep-learning based image segmentation using persistent homology. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 16–28. Springer, 2019.

James R Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew P King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8766–8778, 2020.

GR DeVore, P Falkensammer, MS Sklansky, and LD Platt. Spatio-temporal image correlation (stic): new technology for evaluation of the fetal heart. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 22(4):380–387, 2003.

Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. Spatial warping network for 3D segmentation of the hippocampus in mr images. In *International Conference on*

Medical Image Computing and Computer-Assisted Intervention, pages 284–291. Springer, 2019.

Su-Zhen Dong and Ming Zhu. Utility of fetal cardiac magnetic resonance imaging to assess fetuses with right aortic arch and right ductus arteriosus. *The Journal of Maternal-Fetal & Neonatal Medicine*, 31(12):1627–1631, 2018.

Su-Zhen Dong, Ming Zhu, Hui Ji, Jing-Ya Ren, and Ke Liu. Fetal cardiac mri: a single center experience over 14-years on the potential utility as an adjunct to fetal technically inadequate echocardiography. *Scientific Reports*, 10(1):1–10, 2020.

Michael Ebner, Guotai Wang, Wenqi Li, Michael Aertsen, Premal A Patel, Rosalind Aughwane, Andrew Melbourne, Tom Doel, Steven Dymarkowski, Paolo De Coppi, et al. An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI. *NeuroImage*, 206:116324, 2020.

Irina Grigorescu, Alena Uus, Daan Christiaens, Lucilio Cordero-Grande, Jana Hutter, A David Edwards, Joseph V Hajnal, Marc Modat, and Maria Deprez. Diffusion tensor driven image registration: a deep learning approach. In *International Workshop on Biomedical Image Registration*, pages 131–140. Springer, 2020.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. *Advances in neural information processing systems*, 32, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Bernhard Kainz, Markus Steinberger, Wolfgang Wein, Maria Kuklisova-Murgasova, Christina Malamateniou, Kevin Keraudren, Thomas Torsney-Weir, Mary Rutherford, Paul Aljabar, Joseph V Hajnal, et al. Fast volume reconstruction from motion corrupted stacks of 2d slices. *IEEE transactions on medical imaging*, 34(9):1901–1913, 2015.

Kevin Keraudren, Maria Kuklisova-Murgasova, Vanessa Kyriakopoulou, Christina Malamateniou, Mary A Rutherford, Bernhard Kainz, Joseph V Hajnal, and Daniel Rueckert. Automated fetal brain segmentation from 2d MRI slices for motion correction. *NeuroImage*, 101:633–643, 2014.

Nadieh Khalili, Nikolas Lessmann, Elise Turk, N Claessens, Roel de Heus, Tessel Kolk, Max A Viergever, Manon JNL Benders, and Ivana Išgum. Automatic brain tissue segmentation in fetal MRI using convolutional neural networks. *Magnetic resonance imaging*, 64:77–89, 2019.

Maria Kuklisova-Murgasova, Gerardine Quaghebeur, Mary A Rutherford, Joseph V Hajnal, and Julia A Schnabel. Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Medical image analysis*, 16(8):1550–1564, 2012.

Yiyi Liao, Sarath Kodagoda, Yue Wang, Lei Shi, and Yong Liu. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 2318–2325. IEEE, 2016.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

David FA Lloyd, Kuberan Pushparajah, John M Simpson, Joshua FP Van Amerom, Milou PM Van Poppel, Alexander Schulz, Bernard Kainz, Maria Deprez, Maelene Lohezic, Joanna Allsop, et al. Three-dimensional visualisation of the fetal heart using prenatal MRI with motion-corrected slice-volume registration: a prospective, single-centre cohort study. *The Lancet*, 393(10181):1619–1627, 2019.

Mjaye L Mazwi, David W Brown, Audrey C Marshall, Frank A Pigula, Peter C Laussen, Angelo Polito, David Wypij, and John M Costello. Unplanned reinterventions are associated with postoperative mortality in neonates with critical congenital heart disease. *The Journal of thoracic and cardiovascular surgery*, 145(3):671–677, 2013.

Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G Elmore, and Linda Shapiro. Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 893–901. Springer, 2018.

Shanthi Mendis, Pekka Puska, Bo Norrving, World Health Organization, et al. *Global atlas on cardiovascular disease prevention and control*. World Health Organization, 2011.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Mahesh R Patel, Roman A Klufas, Ronald A Alberico, and Robert R Edelman. Half-fourier acquisition single-shot turbo spin-echo (haste) mr: comparison with fast spin-echo mr in diseases of the brain. *American journal of neuroradiology*, 18(9):1635–1640, 1997.

Kelly Payette, Raimund Kottke, and Andras Jakab. Efficient multi-class fetal brain segmentation in high resolution MRI reconstructions with noisy labels. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pages 295–304. Springer, 2020.

Jialin Peng and Ye Wang. Medical image segmentation with limited supervision: a review of deep network models. *IEEE Access*, 9:36827–36851, 2021.

Esther Puyol-Antón, Bram Ruijsink, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, and Andrew P King. Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 413–423. Springer, 2021.

Paula Ramirez Gilliland, Alena Uus, Milou PM van Poppel, Irina Grigorescu, Johannes K Steinweg, David FA Lloyd, Kuberan Pushparajah, Andrew P King, and Maria Deprez. Automated multi-class fetal cardiac vessel segmentation in aortic arch anomalies using t2-weighted 3d fetal mri. In *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*, pages 82–93. Springer, 2022.

Mina Rezaei, Haojin Yang, and Christoph Meinel. Whole heart and great vessel segmentation with context-aware of generative adversarial networks. In *Bildverarbeitung für die Medizin 2018*, pages 353–358. Springer, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Daniel Rueckert, Luke I Sonoda, Erica RE Denton, S Rankin, Carmel Hayes, Martin O Leach, Derek LG Hill, and David John Hawkes. Comparison and evaluation of rigid and nonrigid registration of breast mr images. In *Medical Imaging 1999: Image Processing*, volume 3661, pages 78–88. International Society for Optics and Photonics, 1999a.

Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999b.

Daniel Salehi, Katrin Fricke, Misha Bhat, Håkan Arheden, Petru Liuba, and Erik Hedström. Utility of fetal cardiovascular magnetic resonance for prenatal diagnosis of complex congenital heart defects. *JAMA network open*, 4(3):e213538–e213538, 2021.

Seyed Sadegh Mohseni Salehi, Seyed Raein Hashemi, Clemente Velasco-Annis, Abdelhakim Ouaalam, Judy A Estroff, Deniz Erdogmus, Simon K Warfield, and Ali Gholipour. Real-time automatic fetal brain extraction in fetal MRI by deep learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 720–724. IEEE, 2018.

Julia A Schnabel, Daniel Rueckert, Marcel Quist, Jane M Blackall, Andy D Castellano-Smith, Thomas Hartkens, Graeme P Penney, Walter A Hall, Haiying Liu, Charles L

Truwit, et al. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 573–581. Springer, 2001.

Richard C Semelka, Nikolaos L Kelekis, David Thomasson, Mark A Brown, and Gerhard A Laub. Haste mr imaging: description of technique and preliminary results in the abdomen. *Journal of Magnetic Resonance Imaging*, 6(4):698–699, 1996.

Matthew Sinclair, Andreas Schuh, Karl Hahn, Kersten Petersen, Ying Bai, James Batten, Michiel Schaap, and Ben Glocker. Atlas-istn: Joint segmentation, registration and atlas construction with image-and-spatial transformer networks. *Medical Image Analysis*, 78: 102383, 2022.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 1013–1020. IEEE, 2018.

Alena Uus, Tong Zhang, Laurence H Jackson, Thomas A Roberts, Mary A Rutherford, Joseph V Hajnal, and Maria Deprez. Deformable slice-to-volume registration for motion correction of fetal body and placenta mri. *IEEE transactions on medical imaging*, 39(9): 2750–2759, 2020.

Alena Uus, Irina Grigorescu, Milou van Poppel, Emer Hughes, Johannes Steinweg, Thomas Roberts, David Lloyd, Kuberan Pushparajah, and Maria Deprez. 3D unet with gan discriminator for robust localisation of the fetal brain and trunk in MRI with partial coverage of the fetal body. *bioRxiv*, 2021.

Alena U. Uus, Irina Grigorescu, Milou P.M. van Poppel, Johannes K. Steinweg, Thomas A. Roberts, Mary A. Rutherford, Joseph V. Hajnal, David F.A. Lloyd, Kuberan Pushparajah, and Maria Deprez. Automated 3d reconstruction of the fetal thorax in the standard atlas space from motion-corrupted mri stacks for 21–36 weeks ga range. *Medical Image Analysis*, 80:102484, 2022a. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841522001311.

Alena U Uus, Milou PM van Poppel, Johannes K Steinweg, Irina Grigorescu, Paula Ramirez Gilliland, Thomas A Roberts, Alexia Egloff Collado, Mary A Rutherford, Joseph V Hajnal, David FA Lloyd, et al. 3d black blood cardiovascular magnetic resonance atlases of congenital aortic arch anomalies and the normal fetal heart: application to automated multi-label segmentation. *Journal of Cardiovascular Magnetic Resonance*, 24(1):1–13, 2022b.

Xiaowei Xu, Tianchen Wang, Yiyu Shi, Haiyun Yuan, Qianjun Jia, Meiping Huang, and Jian Zhuang. Whole heart and great vessel segmentation in congenital heart disease using

deep neural networks and graph matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 477–485. Springer, 2019.

Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. Springer, 2019.

Lequan Yu, Xin Yang, Jing Qin, and Pheng-Ann Heng. 3D fractalnet: dense volumetric segmentation for cardiovascular MRI volumes. In *Reconstruction, segmentation, and analysis of medical images*, pages 103–110. Springer, 2016.

Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31 (3):1116–1128, 2006.

Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8553, 2019.

Appendix A. Alternative segmentation strategies

Here we present two additional approaches to our proposed segmentation strategy. These represent the alternative way of combining our dataset labels (multi-class propagated labels and manually generated binary labels). The strategy behind these new approaches is to learn to split binary labels into multi-class labels. We examine the following:

1. **Attention U-Net Man + Split**: Attention U-Net that splits the predicted CNN binary labels from Attention U-Net Man into multi-class labels.

2. **Attention U-Net Split**: two-step strategy, that involves first splitting the manually segmented binary labels into multi-class (with a CNN), followed by training an additional Attention U-Net using these split labels.

Both these approaches rely on a trained splitting network. This splitting network is trained on joined propagated labels as input (binary) and learns to split these into multi-class, thereby learning to preserve the shape of the input segmentation.

For *Attention U-Net Man + Split*, we use *Attention U-Net Man* to generate the binary labels at inference time, see Fig. 21.
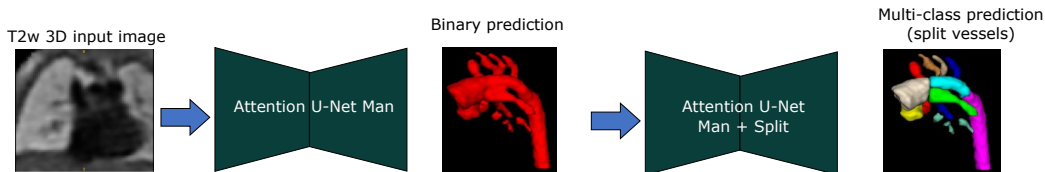
Figure 21: Inference time predictions for *Attention U-Net Man + Split*, which does not require any labels. A binary prediction is generated for the input image, which is then split.

In accordance with our other segmentation experiments, we repeat each experiment three times, to account for network stochasticity.

A.1 Quantitative Results

An important training disadvantage of these two approaches is that they require twice the computational time compared to our proposed method (*Attention U-Net LP + Man*), given that two CNNs are required instead of just one.

Attention U-Net Man + LP and *Attention U-Net Man + Split* are compared in Fig. 22 for the aortic arch (key anomaly area). These results display more extreme results for *Attention U-Net Man + Split*. A Mann-Whitney U test reveals a statistically significant difference (p-value=8.23×10^{-9}) between *Attention U-Net LP + Man* and *Attention U-Net Man + Split*, however not between *Attention U-Net Split* and *Attention U-Net LP + Man*.

A.2 Qualitative Results

The inherent topological issues observed in the binary network (*Attention U-Net Man*) are propagated to its subsequent segmentation splitting variant, *Attention U-Net Man + Split*. This causes vessel misclassification, as displayed in two independent test set cases in Fig. 23. Undetected small vessels are also signalled.

(a) ASD scores (lower = better).

(b) Dice scores (higher = better).

Figure 22: Quantitative scores for the aortic arch (anomaly area) on test set comparing our proposed labelling type training combination (*Attention U-Net LP + Man*) to *Attention U-Net Man + Split*. We repeat each experiment three times and include all results here.

Attention U-Net Man + Split takes the output of *Attention U-Net Man* as input and performs vessel splitting, learning to preserve the shape of the input segmentation. This leads to the perpetuation of the existing topological problems in the predicted segmentation. Moreover, when vessel merging occurs, the splitting network further hampers performance by misclassifying merged vessels. This vessel misclassification issue is not as persistent in *Attention U-Net Split*, the network which is trained on split manual labels, however still occurs on lower quality cases, particularly cases presenting challenging discernment between aorta and MPA.

We quantify this by conducting a qualitative assessment of the segmentation topology area (see Section 3.4.2 for definition). We include scores for all vessels and for the anomaly area in Fig. 24. This showcases the superior performance of our proposed method, followed by *Attention U-Net Split*.

Appendix B. Anomaly classifier

Figs. 25a, 25b, 25c, 25e and 25d contain our classifier test set confusion matrices for our three training rounds.

B.1 Qualitative analysis extended

Here we extend our investigation into the lower-scoring subjects of our final framework (*Attention U-Net LP + Man + Classifier* with **DenseMulti** classifier), as assessed by our qualitative analysis (Sec. 4.4). We include our topology scores separated for CoA and RAA in Figs. 26.

In total, we find ten nine subjects with a topology score of 2 for the anomaly area, and one with a score of 3. These subjects comprise our two misclassified RAA cases (due to a partially and fully segmented double arch), as well as three CoA cases repeated across all
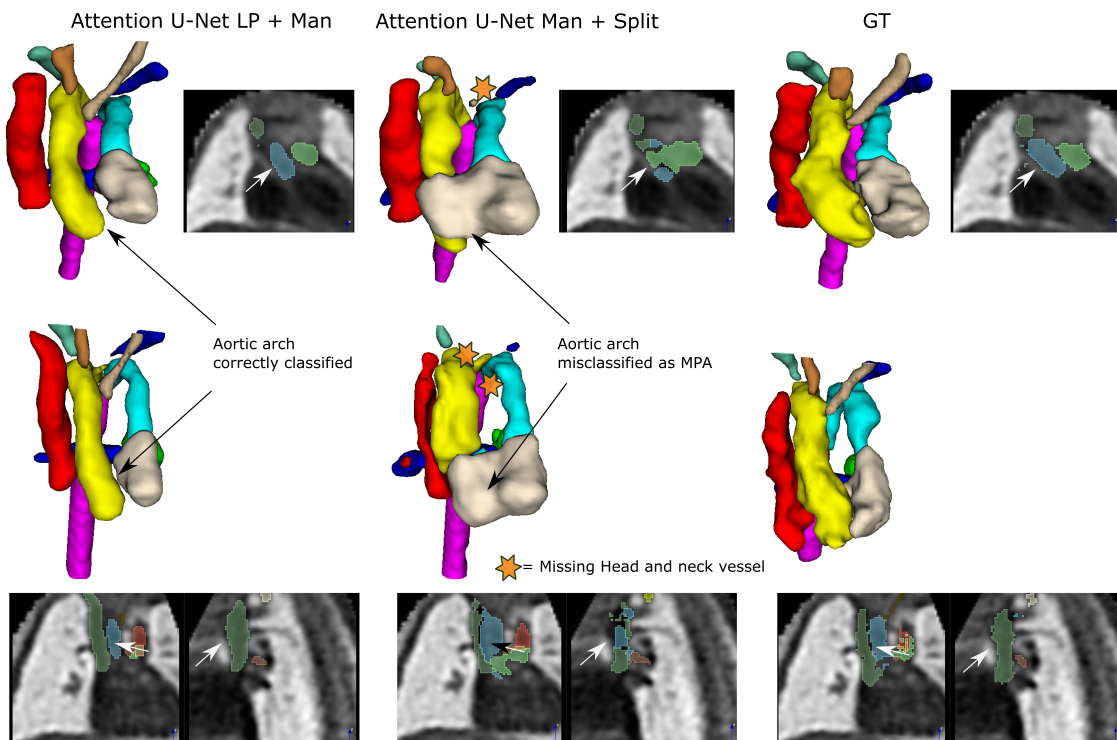
Figure 23: Vessel misclassification issue observed in *Attention U-Net Man + Split*, with comparison to our proposed labelling combination *Attention U-Net LP + Man* and manually segmented GT. Predictions are shown for two independent test set cases. Orange stars indicate missing vessels in the splitting network, and white arrows (T2w images) indicate the misclassification area.

three rounds. The latter cases include oversegmentation of the aorta into head and neck vessels, and merging into AD. We inspect the images and find lower quality and alignment compared to the atlas. Therefore we conclude that these subjects are outliers due to lower image quality, which is supported by the fact that three independently trained networks present segmentation issues on these subjects. Fig. 27 showcases the three CoA cases with a score of 2, alongside a high-quality correctly segmented case, and our CoA atlas to highlight the image quality differences.

B.2 Latent space representations

Fig. 28 contains our t-SNE reduced latent space visualisations before and after adding a classifier to our segmentation frameworks, for one of our training rounds. We display both our binary segmentation networks (*Attention U-Net Man*), and our proposed multi-class approach (*Attention U-Net LP + Man*). We observe slightly further clustering after joint training with the anomaly classifier. However, this is not as pronounced as training using multi-class labels as opposed to binary manually segmented labels.
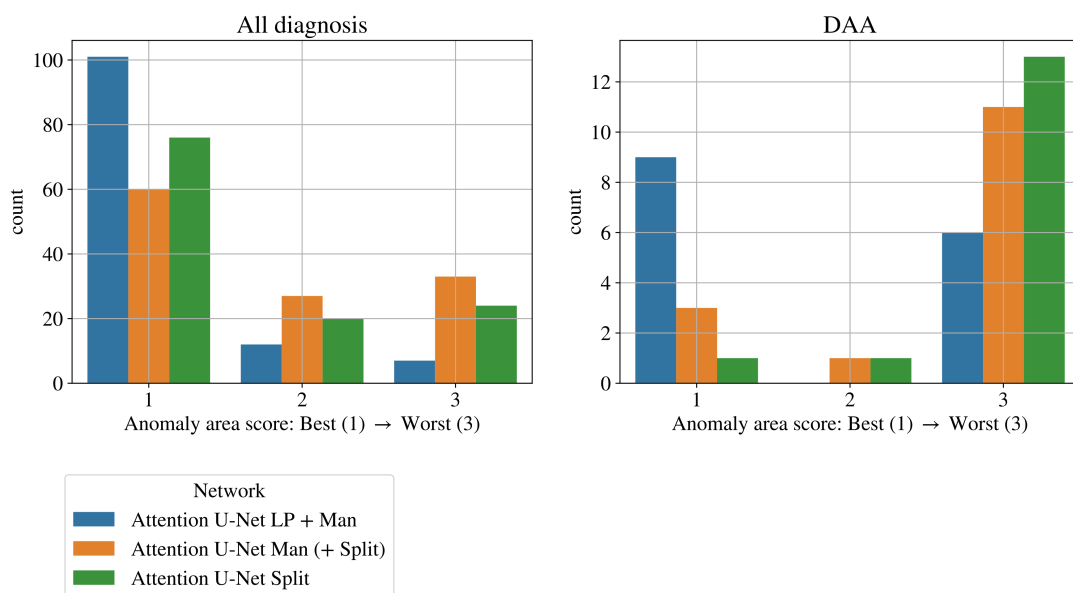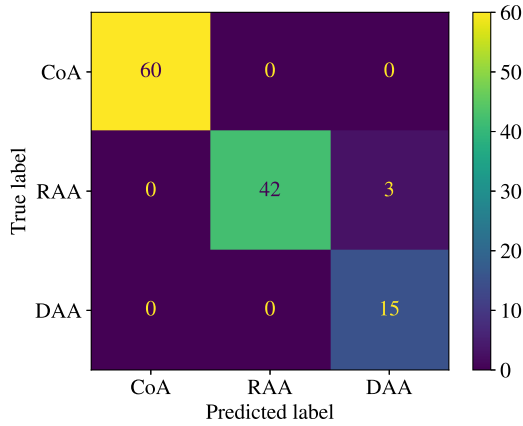
Figure 24: Qualitative analysis results on topology correctness of the anomaly area (aortic arch) on our test set for **all diagnoses** (left), and for DAA cases only (right). Each training experiment was repeated three times, and we include all of our results here (i.e. we have three results for each of our 40 test set cases, one for each round, so the total number of cases is 120)
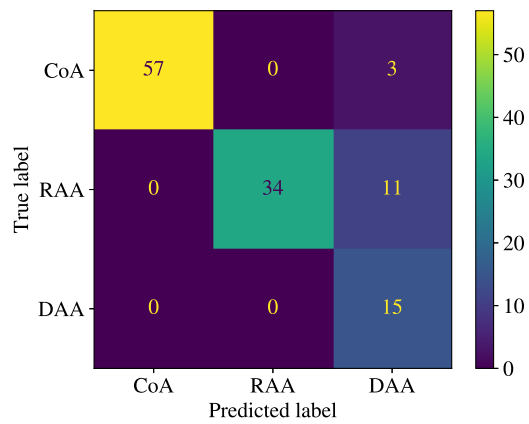
.

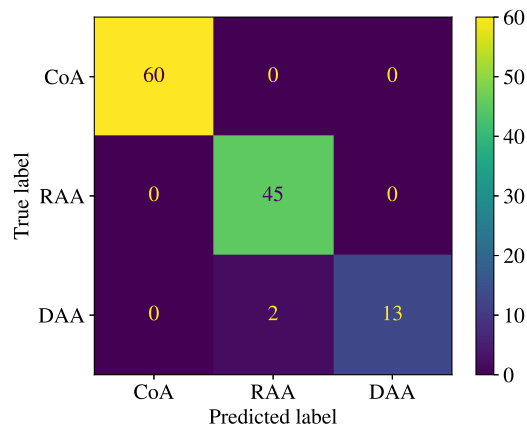(a) **DenseMulti** confusion matrix, three training rounds (1 misclassification per round).
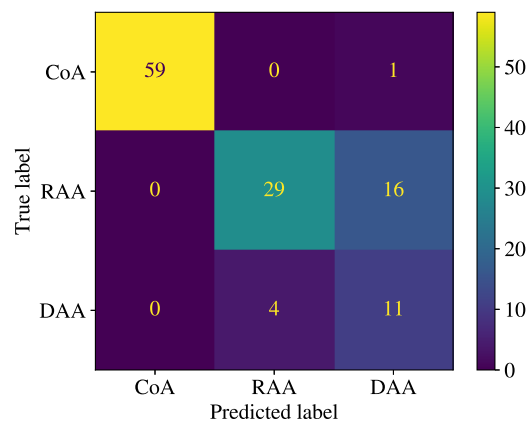
(b) **DenseImage** confusion matrix, three training rounds.
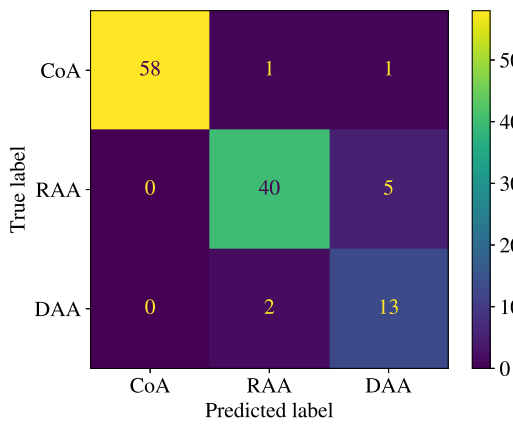
(c) **DenseBin** confusion matrix, three training rounds.

(d) **DenseBin** before joint training confusion matrix, three training rounds.

(e) **DenseMulti** before joint training confusion matrix, three training rounds.
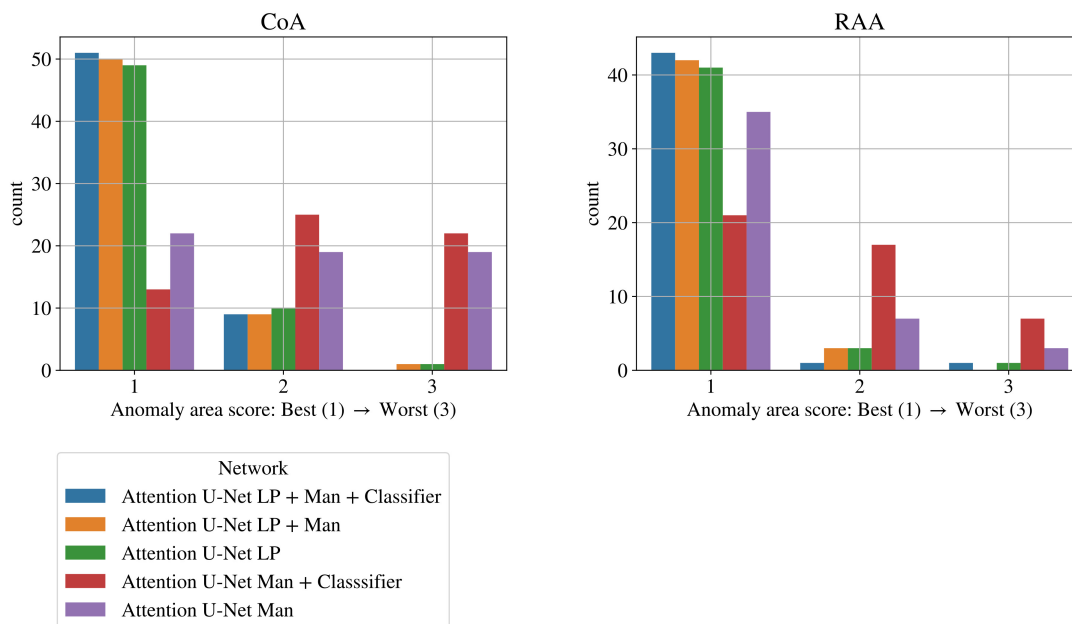
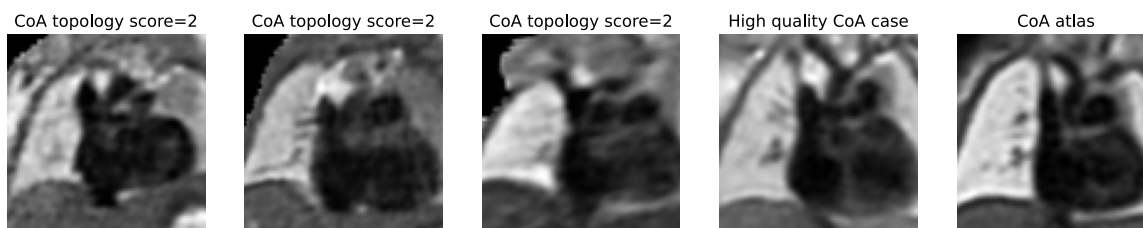Figure 26: CoA (left) and RAA (right) test set topology scores.

Figure 27: CoA test set cases with a topology score of 2 in our final framework, alongside a high-quality example case and our CoA atlas.
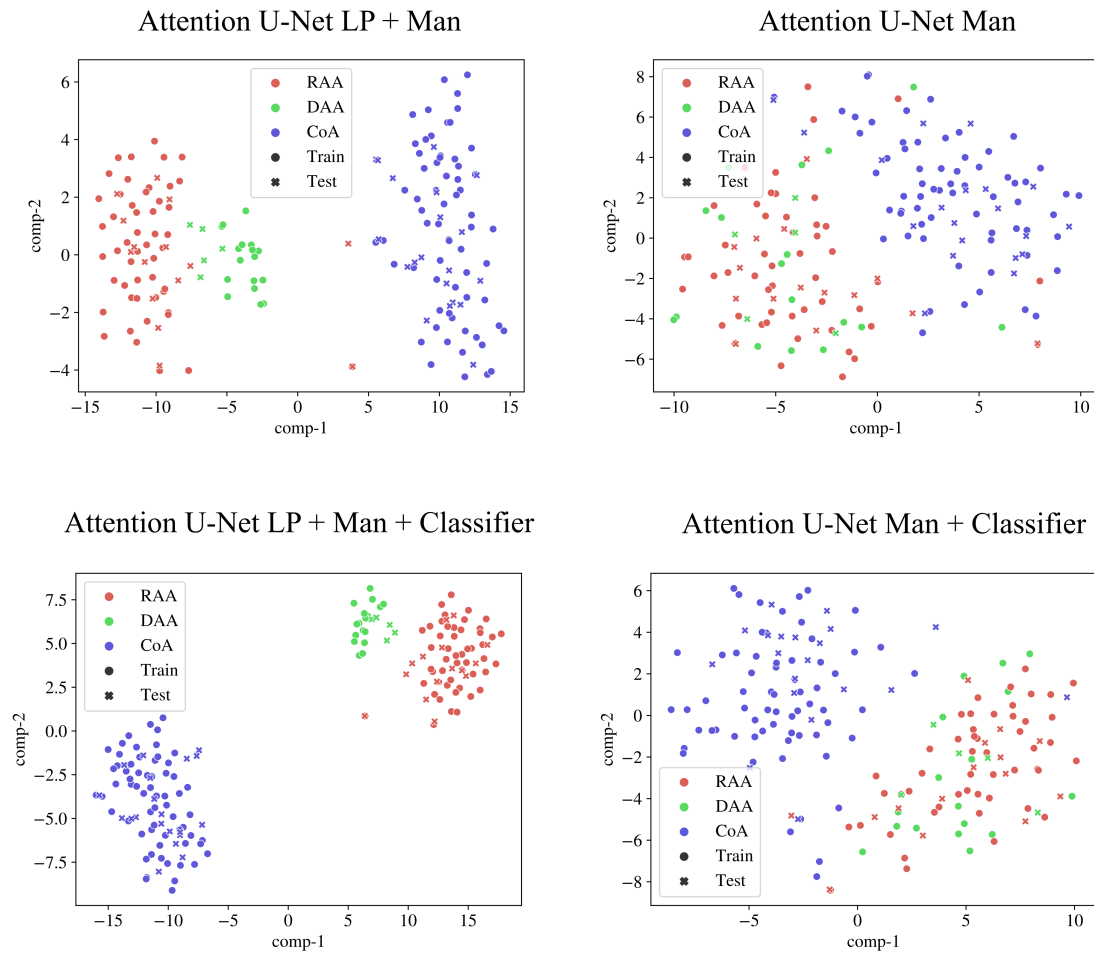
Figure 28: t-SNE visualisation of our network bottleneck features before (top row) and after (bottom row) the addition of an anomaly classifier. This is an example of one of our training rounds.