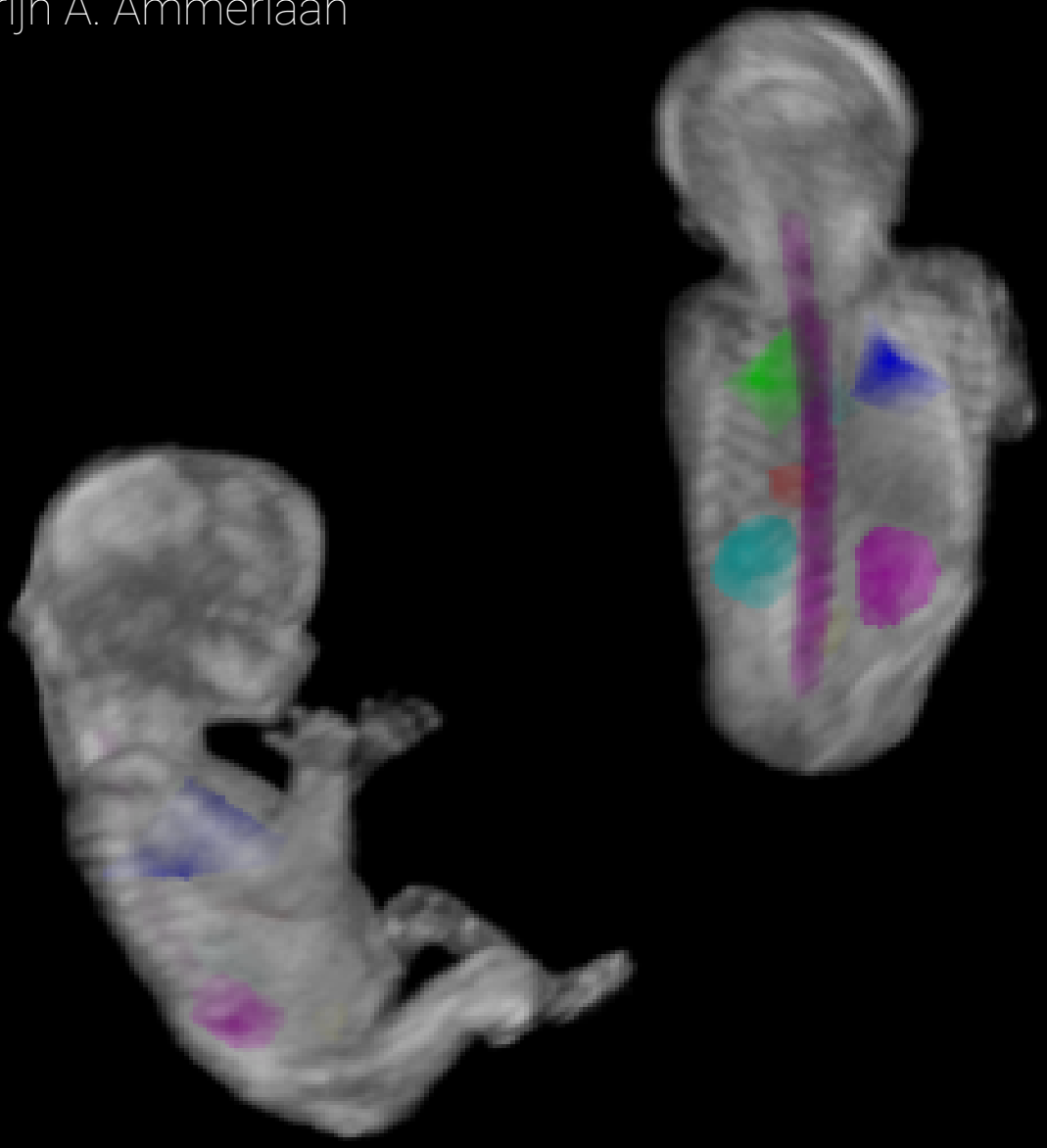
Master Thesis

Automated organ detection for first-trimester anomaly screening in three-dimensional fetal ultrasound

Rozemarijn A. Ammerlaan



Erasmus MC

TU Delft

Automated organ detection for first-trimester anomaly screening in three-dimensional fetal ultrasound.

by

Rozemarijn A. Ammerlaan

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday November 14, 2025 at 13:00.

Student number:	4830571	
Project duration:	March 1, 2025 – November 14, 2025	
Thesis committee:	Dr. ir. F.M. Vos,	TU Delft
	ir. M.C. Zijta,	Erasmus MC
	Dr. ir. W.A.P. Bastiaansen,	Erasmus MC
	Dr. ir. A.H.J. Koning,	Erasmus MC

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft Delft University of Technology

Erasmus MC Universitair Medisch Centrum Rotterdam

Abstract

To identify congenital anomalies, pregnant women are offered a structural screening using ultrasound during the first trimester. This screening is performed in two-dimensional ultrasound and has a duration of up to 45 minutes. While three-dimensional (3D) and 3D virtual reality (VR) ultrasound have shown to be beneficial for the anomaly detection rate of some fetal structures, their implementation is limited by an increased evaluation time and the need for expert knowledge. To address these problems and reduce operator dependency, an automatic organ detection model for first-trimester fetal ultrasound is proposed. We used 69 3D ultrasound scans from the VR-FETUS study with a gestational age ranging from 11+0 weeks and 13+6 weeks. These scans were annotated in VR with sparse labels (landmarks) for the heart, lungs, spine, choroid plexuses, cerebellum, mandible, orbits, upper lip and nasal bone, while dense labels (segmentations) were provided for the bladder, kidneys and stomach. In total 50 scans were used for training and 19 scans were used for testing. We trained nnU-Net models for each organ separately, using pseudo segmentation labels that were created from the landmarks. Furthermore, it was tested if using a combination of labels, additional training data or different loss functions would increase model performance. The Dice similarity coefficient (DSC) was used for evaluation of the segmentation labels and we assessed the detection rate for all models. The heart, lungs and mandible were detected in 95 to 100% of the test scans, while the cerebellum and plexuses were detected in 63 tot 79% of the test scans. The detection rate for the orbits and upper lip was lower than 43% and the nasal bone was not detected by any model. The median DSC for the bladder, kidneys and stomach ranged between 0.70 and 0.81 and they were detected in 89 to 100% of the test scans. With our results we take the first step towards automatic organ detection in first-trimester 3D ultrasound, by lowering the evaluation time and reducing the operator dependency.

1. Introduction

Worldwide, approximately 6% of babies are born with a congenital disorder, which can consist of functional or structural anomalies [1]. These anomalies may result from gene defects, chromosomal disorders or nutrient deficiencies [1]. In some cases no cause can be identified. To screen for anomalies and assess overall fetal development, pregnant women are offered a non-invasive prenatal test (NIPT) and a structural ultrasound scan in the first trimester. The NIPT is used for screening of the maternal blood for chromosomal anomalies [2]. Alternatively, ultrasound allows screening for anomalies that are not detectable by the NIPT. The first trimester anomaly screening, which was introduced in 2021, is still under evaluation [3]. Early detection of anomalies allows more time for counseling and decision making [4].

During the first-trimester ultrasound examination, a sonographer follows the guidelines of the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG). This examination should be performed between 11+0 and 13+6 weeks of gestational age (GA) [5]. It includes checking the heartbeat of the fetus, confirming an intrauterine pregnancy, measuring fetal biometrics, such as crown-rump length (CRL) and biparietal diameter (BPD); and assessing the fetal anatomy [5]. For the anatomical assessment, eight different anatomical regions should be checked. These regions are examined using multiple two-dimensional (2D) ultrasound viewing planes, leading to a total scan duration of up to 45 minutes [4]. However, if the fetus is not positioned in the desired orientation, the scan could take even longer, which may lead to incomplete checking or all anatomical regions.

The current ISUOG guidelines are designed for 2D ultrasound scans. However, the implementation of three dimensional (3D) ultrasound scans has demonstrated advantages over 2D scans, including reduced scanning time and improved anomaly detection [6, 7]. Additionally, a study showed that the use of virtual reality (VR) for evaluation of 3D ultrasound scans resulted in an enhanced depth perception and a more intuitive visualization of the fetus [8]. Furthermore, the study also found that the use of 3D VR ultrasound resulted in an increased sensitivity for specific abnormalities, particularly in the skeleton and extremities. However, expertise in VR was required and the evaluation time

using 3D VR was higher compared to evaluation solely with 3D ultrasound.

As a large part of the evaluation time is taken up by the anatomical assessment, automatic organ detection methods have been explored in literature. The implementation of an automated organ detection model could potentially reduce the evaluation time and minimize the operator dependent variability across scans. Although various fetal organ detection models are available [9, 10, 11, 12], none were trained for multi-organ detection of the fetus in first-trimester 3D ultrasound. Therefore, the goal of this study is to train a deep learning model for multi-organ detection in 3D first-trimester fetal ultrasound volumes. This model could assist the sonographer by highlighting the location of specific organs, as described in the ISUOG guidelines. The goal is to be able to correctly identify and locate multiple organs within 3D ultrasound volumes. This is a particularly complex task because it involves the detection of a large number of organs, whereas most previous work focused on only a few organs. Additionally, we have to deal with the challenges inherent to ultrasound, such as low resolution, noise and different types of artifacts [13]. To address these difficulties, we trained several deep learning models during a series of experiments, to determine what configurations resulted in the best performances. To the best of our knowledge, this study presents the first explorations into multi-organ detection using first-trimester 3D VR fetal ultrasound data.

2. Methods and materials

2.1. Approach

Organ detection refers to identifying the presence and location of an organ in an image, while organ segmentation involves the precise labeling of a subset of pixels or voxels in an image, including the coverage of the boundaries and internal volume [14, 15]. Given that segmentation also requires localization of the organ, segmentation can be considered as an extension of detection. Although the primary objective of this project is organ detection, we will also look into the potential of organ segmentation, which is facilitated by the availability of a heterogeneous dataset, containing both sparse and dense annotations.

Due to its state-of-the-art performances and generalization across medical imaging data, the 'no new' U-Net framework was chosen as segmentation model [16]. Previous work using nnU-net for segmentation of the fetal head volume and embryonic volume demonstrated promising performances [17]. A U-Net follows an encoder-decoder structure with skip connections [18]. These skip connections allow the model to capture both global and fine details, enabling it to learn complex anatomical shapes from limited input data. nnU-Net is a full pipeline for medical image segmentation build around a U-Net architecture. It includes preprocessing, data augmentation, post-processing and network design, all based on dataset characteristics, such as image size and intensity distribution. The network input should consist of a set of original input images with (multi-class) segmentation masks and the output consists of predicted segmentation masks for the original input images.

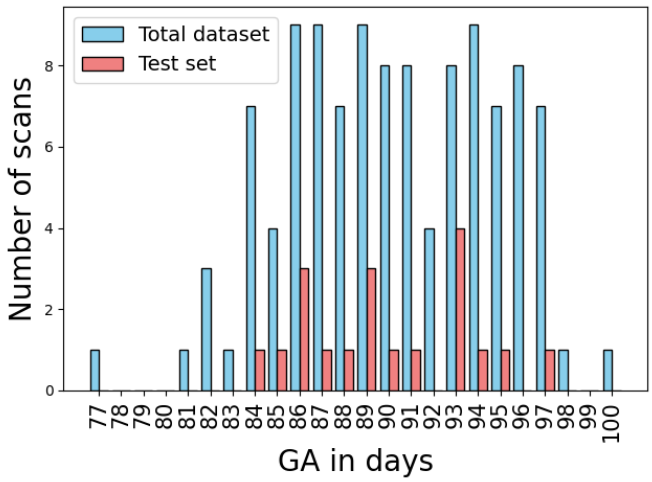


Figure 1: *Histogram showing the distribution of the number of scans for every gestational age (GA) in days of the total dataset (blue) and the test set (pink).*

2.2. Data

2.2.1 Dataset description

The dataset used in this project is part of the VR-FETUS study [4]. This study is a randomized trial that was performed at the gynecology department of the Erasmus Medical Center during the period between June 2017 and September 2021. Women were included if they were 18 years or older, had a pregnancy duration between 11+0 and 13+6 weeks of GA and if they were referred for a high risk pregnancy. Women were ex-

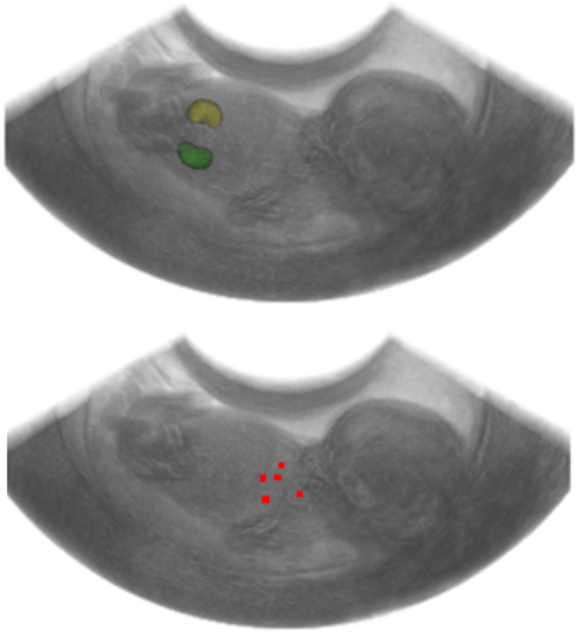


Figure 2: *Examples of annotation labels for different organs. Top: segmentation labels for the right and left kidney. Bottom: landmarks for the heart.*

cluded if the pregnancy was terminated before the ultrasound scan or if an anomaly was found before randomization. Full anatomical screening of the fetus was performed, during which 3D ultrasound volumes of the fetal head and body were acquired additional to the standard 2D planes. Scanning was performed using a high frequency transvaginal transducer (frequency of 4-9 or 6-13 megahertz (MHz)) and a lower frequency transabdominal transducer (2-6 MHz), both using a Voluson E10 machine (GE Healthcare). Scan and voxel dimensions varied between scans, voxel spacings within each scan were isotropic. Fetal orientation differed between scans, due to subject variations and positional changes of the fetus. Fully annotated scans were available for 69 subjects. Additional scans were annotated with the bladder and kidneys only. An overview of the distribution of gestational ages of the subjects is shown in Figure 1.

2.2.2 Annotations

The ultrasound scans were annotated in 3D VR using a standardized protocol, both by an experienced fetal medicine specialist and a medical student. The labels consisted of landmarks placed around the regions of interest of 9 different organs and 3D segmentations of the bladder, kidneys and stomach. The annotated organs were selected based on the ISUOG guidelines for first trimester screening [5]. Example annotations are shown in 3D in Figure 2 and in 2D in Figure 4. The labels were created using the V-Scope software, which creates holograms of the ultrasound volumes in VR and has tools for annotating the volumes [19].

The bladder, heart, kidneys, lungs, spine and stomach were annotated in the fetal body volume. The cerebellum, choroid plexuses, mandible, nasal bone, orbits and upper lip were annotated in the fetal head volume. It should be noted that not all annotations for a subject were always created in the same ultrasound volume. This is because not all organs were always visible in the same volume due to ultrasound artifacts or low image quality. These missing organs were then annotated in a different volume of the same subject.

2.3. Mask creation based on landmarks

The expected input labels for training the model consist of segmentation masks, which were available for the bladder, kidneys and the stomach. For the remaining organs, annotated only with landmarks, different types of pseudo labels were created. For the heart, lungs and choroid plexuses convex hulls were created, encompassing the landmarks. This ensured that the pseudo label was located within the boundaries of the organ. However, the convex hull did not fully capture the anatomical boundaries. An example of the creation of a convex hull is shown in Figure 3A.

For anatomically elongated or curved structures, such as the spine and mandible, pseudo labels were created by connecting the landmarks with small cylinders. An example of this is shown in Fig 3C. The radius of these cylinders was empirically chosen based on the type of organ, for the spine the radius was 4 voxels, for the mandible and cerebellum 3 voxels and for the upper lip 2 voxels. The connected cylinders covered most of the original structure. However, they did not exactly follow the anatomical boundaries. The pseudo labels for the orbits were created by fitting a small sphere in between the two landmarks for each orbit. The diameter of the sphere was chosen as the distance between two landmarks for one orbit, an example can be found in Fig 3C. 2D examples of the created pseudo labels are shown in Figure 4.

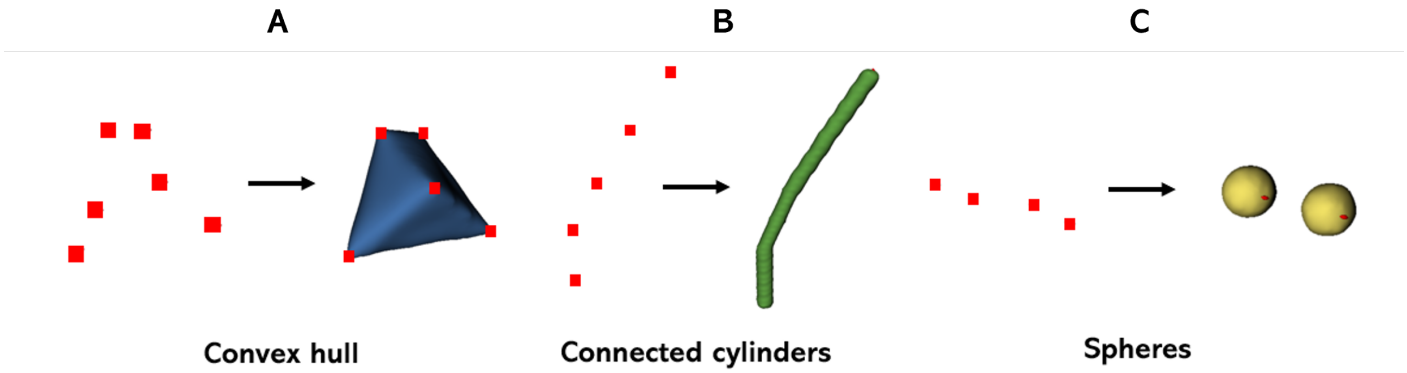


Figure 3: *Examples of the creation of pseudo labels from the landmarks (red) using a convex hull (A), connected cylinders (B) and spheres (C).*

2.4. Loss functions

Several loss functions were used during training of the models, including individual and combinations of loss functions.

Soft Dice loss

The multi-class soft Dice loss (L_{SD}):

$$\mathcal{L}_{SD}(y_c, \hat{y}_c) = 1 - \frac{1}{C} \sum \frac{2TP_c + \epsilon}{2TP_c + FP_c + FN_c + \epsilon} \tag{1}$$

$$\begin{aligned} TP_c &= \sum(\hat{y}_c * y_c) \\ FP_c &= \sum(\hat{y}_c * (1 - y_c)) \\ FN_c &= \sum((1 - \hat{y}_c) * y_c) \end{aligned} \tag{2}$$

is based on the overlap of the ground truth and predicted labels [20]. The loss is averaged over the classes C and calculated using the true positives TP_c, true negatives TN_c and false positives FP_c using the ground truth (y_c) and predicted labels ($\hat{y}_c$), see Equation 2. The small constant ϵ (set to 1×10^{-5}) is added for numerical stability. A soft Dice loss of 1 indicates no overlap of the prediction and ground truth, a soft Dice loss of 0 indicates full overlap.

Cross-Entropy loss

The cross-entropy loss ($\mathcal{L}_{CE}$):

$$\mathcal{L}_{CE}(y_{i,c}, \hat{y}_{i,c}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} log(\hat{y}_{i,c}) \tag{3}$$

is a per-voxel based loss function that calculates the differences between the ground truth labels $y_{i,c}$ and the predicted class probabilities $\hat{y}_{i,c}$ [20]. It is calculated over C classes and a total of N voxels in the image. Minimization of the $\mathcal{L}_{CE}$ loss aims to get $\hat{y}_{i,c}$ close to 1 for correct classes and close to 0 for wrong classes.

Focal loss

The focal loss function ($\mathcal{L}_F$):

$$\mathcal{L}_F(\hat{y}_t) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \alpha_t (1 - \hat{y}_t)^{\gamma} log(\hat{y}_t) \tag{4}$$

$$\hat{y}_t = \begin{cases} \hat{y}, & \text{correctly predicted} \\ 1 - \hat{y}, & \text{not correctly predicted} \end{cases} \tag{5}$$

prioritizes the misclassified or uncertain voxels, instead of weighing all voxels equally [20]. This makes the focal loss suitable for medical images, which frequently present an unbalance between the foreground classes and the background class. The multi-class focal loss is calculated over all voxels N and all classes C. α_t can be chosen such that the loss is higher for misclassifications of the foreground class. γ is the focusing parameter, deciding how much should be focused on the voxels

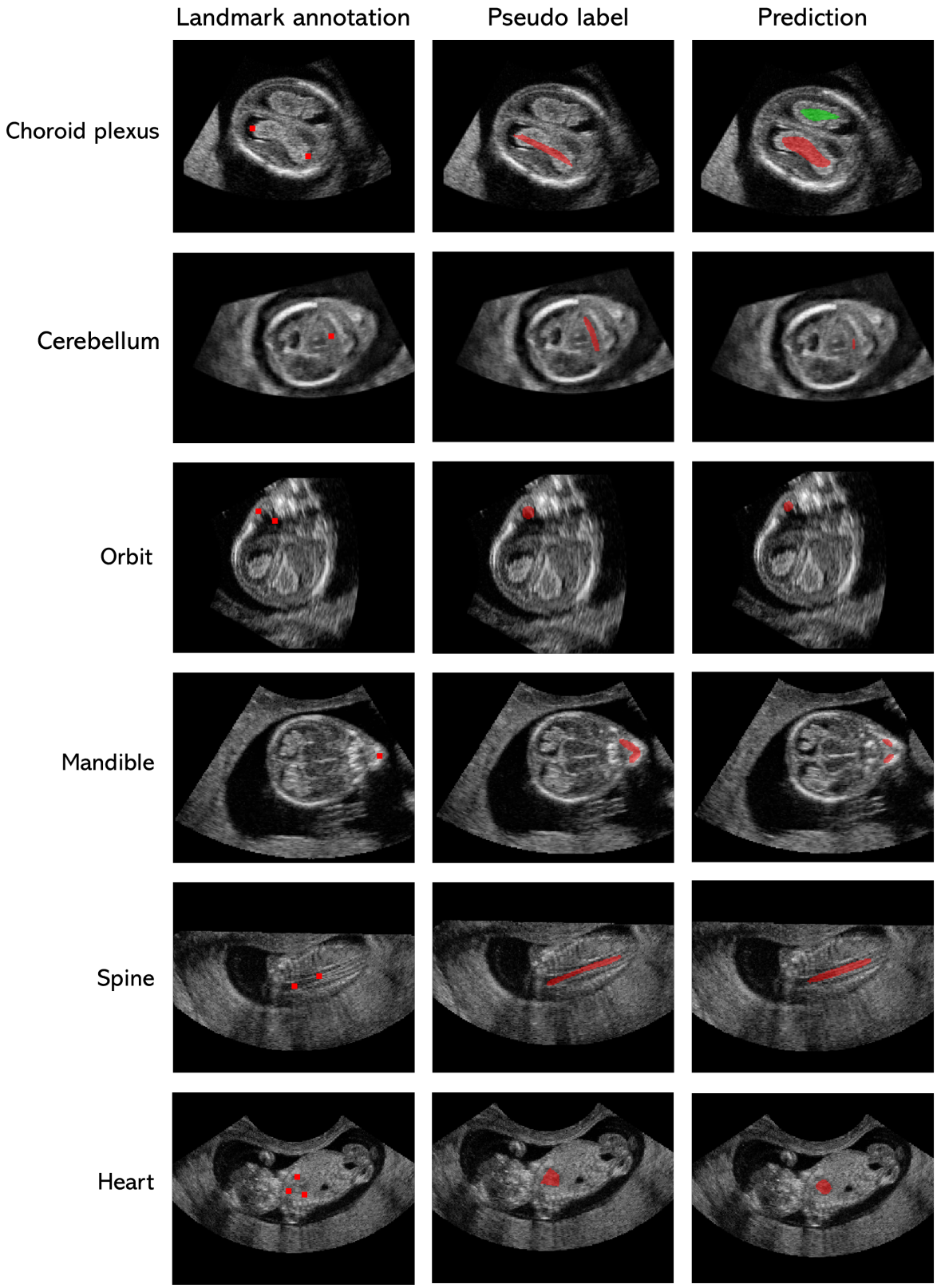


Figure 4: *Examples of the annotated landmarks, created pseudo labels and predictions by the trained models for the choroid plexus, cerebellum, orbit, mandible, spine and heart. The labels are shown in red in two-dimensional ultrasound images.*

that are difficult to classify. The probability that a voxel belongs to the foreground, is stated as $\hat{y}$, see Equation 5. If the predicted probability of a voxel belonging to the foreground ($y = 1$) is low, the focal loss will be high.

Centroid distance loss

The centroid distance loss ($\mathcal{L}_{\text{CD}}$):

$$\mathcal{L}_{\text{CD}}(\vec{m}_y, \vec{m}_{\hat{y}}) = \frac{1}{C} \sum_{c=1}^{C} \left\| \vec{m}_{\hat{y}} - \vec{m}_y \right\|_2 \tag{6}$$

is a function that aims to minimize the distance between the centroids of the prediction and ground truth labels [21]. The centroid distance loss can be considered for segmentation problems for which the overlap is less important than the correct location of the centroid of the mask. First the centroids of the ground truth $\vec{m}_y = [\vec{m}_y^1, \ldots, \vec{m}_y^C]$ and centroids of the prediction $\vec{m}_{\hat{y}} = [\vec{m}_{\hat{y}}^1, \ldots, \vec{m}_{\hat{y}}^C]$ are calculated by finding the center of mass of the masks for each class. Then the multi-class centroid distance loss is calculated by computing the Euclidean distance for all classes C, as in Equation 6.

2.5. Training characteristics

Training of the nnU-Net models for organ detection was done on a Snellius NVIDIA A100 GPU, with up to 80GB of memory. All models were trained for 150 epochs for each fold, as no improvement of the validation Dice similarity coefficient, training loss and validation loss was observed after 150 epochs.

The stochastic gradient descent optimizer was used in training, with Nesterov momentum and a poly learning rate scheduler. This allows for a non-linear decrease of the learning rate. Batch size for training was 2, with varying patch sizes depending on the image size. During training, 5-fold cross validation was applied.

2.6. Dataset split

For training and analyzing the trained nnU-Net models, the dataset was split into train and test sets. Since not all organs were annotated within the same volumes for every subject, it was chosen to partition the dataset by subject rather than volume. The test set was selected based on two criteria: for each subject a complete set of organ labels should be available and the gestational ages of the subjects should reflect the distribution of gestational ages within the dataset. The final test set consisted of a total of 19 subjects. The amount of available labels varied between the organs, causing different numbers of training samples per trained model. An overview of the number of train and test samples for each organ is shown in Table 1

Table 1: *Overview of the number of subjects for the train and test sets for different organs.*

Organ	Training set	Test set
Heart	47	19
Lungs	47	19
Spine	44	19
Stomach	47	19
Kidneys	50	19
Bladder	46	19
Choroid plexuses	44	19
Cerebellum	41	19
Mandible	44	19
Orbits	40	19
Upper lip	40	19
Nasal bone	41	19

2.7. Evaluation metrics

The performances of our models were evaluated using different evaluation metrics. For evaluation of organs that were annotated with segmentation masks, the Dice similarity coefficient (DSC) was used [20]. The DSC is the complement of the Dice loss:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{7}$$

The formulas for the TP, FP and FN can be found in Equation 2. A score of 1 means full overlap and a score of 0 means no overlap.

For evaluation of the organs annotated with landmarks, the Dice similarity coefficient was not an appropriate metric, since no ground truth segmentation masks were available for comparison. For these organs, a centroid distance metric (CDM) was used for evaluation. The centroid of both the prediction $\vec{m}_{\hat{y}}$ and ground truth $\vec{m}_{\hat{y}}$ mask was calculated and the Euclidean distance was computed as in Equation 6. The CDM was normalized for each organ by using the characteristic length of the volume of the ground truth segmentation mask:

$$CDM_{norm} = \frac{CDM}{\sqrt[3]{V_{organ}}} \tag{8}$$

The CDM was not a representative metric for elongated structures such as the spine, as it did not take the spread of the boundary points into account. Therefore, also the 95th percentile Hausdorff distance ($HD95$) was calculated for all organs:

$$\begin{aligned} d_{95}(A,B) &= x_{95^{\text{th}}}\Big(\{\min_{b\in B} d(a,b)\}_{a\in A}\Big), \\ HD95(A,B) &= \max\big\{d_{95}(A,B), d_{95}(B,A)\big\}. \end{aligned} \tag{9}$$

$$HD95_{norm} = \frac{HD95}{\sqrt[3]{V_{organ}}} \tag{10}$$

The $HD95$ measures the 95th percentile of the maximum of the minimum distances $d(a,b)$ and $d(b,a)$ between the sets of boundary points A, B of the ground truth and prediction mask, excluding outliers. The $HD95$ was normalized for the size of each organ as well, see Equation 10.

For all organs, the number of times that the model could correctly predict the organ of interest was determined as a percentage of the total test scans, the detection rate (DR). The criterion for detection was: the ground truth and prediction masks should overlap ($DSC > 0$).

2.8. Statistical analysis

To assess whether experiments significantly outperformed each other based on the metrics, the Wilcoxon signed-rank test was used. For the test, p-values were only calculated for organs that were detected by both models in the comparison (paired samples). A p-value of < 0.05 was seen as significant. Bootstrapping was applied for calculation of interquartile range of the DSC, CDM_{norm} and $HD95_{norm}$ values. This was done by resampling the metrics for 1000 iterations (with replacement) and recalculating the evaluation metrics for each resampled subset. The metrics for all detected organs in the test scans were taken into account for bootstrapping.

3. Experimental setup

3.1. Experiment 1: (Pseudo) labels

In this experiment, a separate model was trained for each organ using the different (pseudo) labels:

1. **Segmentation label:** stomach, bladder, kidneys
2. **Convex hull:** heart, lungs, choroid plexuses
3. **Connected cylinders:** spine, mandible, cerebellum, upper lip, nasal bone
4. **Spheres:** orbits

The aim of this experiment was to develop single organ models, serving as a reference for comparative analysis. The model performances were evaluated using DSC for the ground truth segmentation labels and the CDM_{norm} and $HD95_{norm}$ for the pseudo labels.

3.2. Experiment 2: Combination of labels

The pseudo labels of different organs were combined into the fetal body and head volume as input for the model. Hypothetically, a combination of organs will yield better results, as the model will have more spatial context of the organs within the ultrasound scan. Two different combinations of input labels were used:

1. **Fetus:** heart, left and right lung, spine
2. **Head:** left and right choroid plexus, left and right orbit, cerebellum, mandible, nasal bone, upper lip

Both models were evaluated using the CDM_{norm} and $HD95_{norm}$ and were compared to the results of the single organ models.

3.3. Experiment 3: Missing labels

The bladder, kidneys and stomach labels were not always annotated in the same volume in the provided dataset. When the model input is missing some of the labels, the model may predict the presence of an organ, while no ground truth label is available, potentially leading to lower model performances. To account for the missing labels, regions that potentially contain specific organs can be labeled with 'ignore', excluding them from training. These voxels are then ignored during training and will not penalize the model. The aim of this experiment is to train a model for multi-label segmentation of the bladder, kidneys and stomach.

For creating the regions that should be ignored during training, the stomach and kidneys were predicted in the volumes containing the bladder, by using the single organ models of Experiment 1. Finally the performance of the combined model with missing labels was compared to the performances of the single organ models of the bladder, kidneys and stomach using the DSC.

3.4. Experiment 4: Dataset size

For the bladder and kidney segmentation labels an additional number of scans was available: 46 for the bladder, 50 for the kidneys. The aim of this experiment was to determine if a larger training set would improve the performances of the model. The models were compared to the single organ models of the bladder and kidneys using the DSC.

3.5. Experiment 5: Loss functions

An evenly weighted combination of the soft Dice loss and binary cross-entropy loss was used for training in all previously described experiments, as it is the standard loss function in the nnU-Net framework. While these losses generally work well for segmentation problems with ground truth segmentation masks available, they might not be suitable for the generated pseudo labels. For that reason, multiple loss functions were tested. For the heart, with a convex hull as pseudo label, the centroid distance loss ($\mathcal{L}_{CD}$) was tested both separately and in combination with the soft Dice loss ($\mathcal{L}_{CD} + \mathcal{L}_{SD}$). The hypothesis was that the centroid distance loss might emphasize the minimization of the centroid distance without optimizing the overlap of the prediction and convex hull. For the nasal bone the focal loss ($\mathcal{L}_F$) was tested, also in combination with the soft Dice loss ($\mathcal{L}_F + \mathcal{L}_{SD}$). Here it was hypothesized that the focal loss might help with detecting smaller structures, as it focuses more on voxels that are hard to classify.

Again, the performances of the models trained with the loss functions were compared to the performances of the single organ models using the CDM_{norm} and $HD95_{norm}$.

4. Results

4.1. Experiment 1: (Pseudo) labels

For this experiment, 12 single organ models were trained using (pseudo) labels. Figure 5 shows an example of the ground truth pseudo labels for the lungs and the predicted segmentation masks by the model. The single organ models correctly predicted 89 to 100% of the test scans for the heart, lungs, spine, kidneys, bladder, stomach and mandible, see Figure 6. For the plexuses, cerebellum, orbits and upper lip, the detection rate was less then 84%. This problem with detection for these single organ models also shows in the box plots of the CDM_{norm} and $HD95_{norm}$ in Figure 8. Both the median CDM_{norm} and $HD95_{norm}$ were above 0.5 for the left and right plexus, the cerebellum, right orbit and the upper lip. During visual inspection of the predictions of the orbits and plexuses, it was observed that the models occasionally predicted the left and right labels for a single side, see Figure 7. For the models that could correctly predict the organs, the median CDM_{norm} and $HD95_{norm}$ were below or close to 0.25.

Figure 9 shows box plots of the DSC for the bladder, left and right kidney and stomach, with median DSC

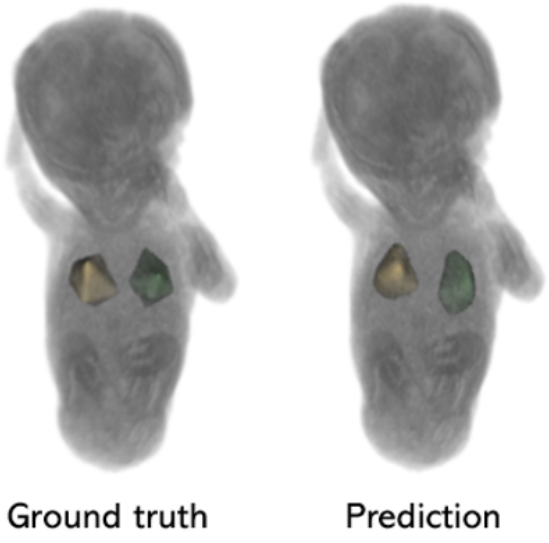


Figure 5: *Example of the ground truth input (left) and prediction (right) for the single organ model of the lungs.*

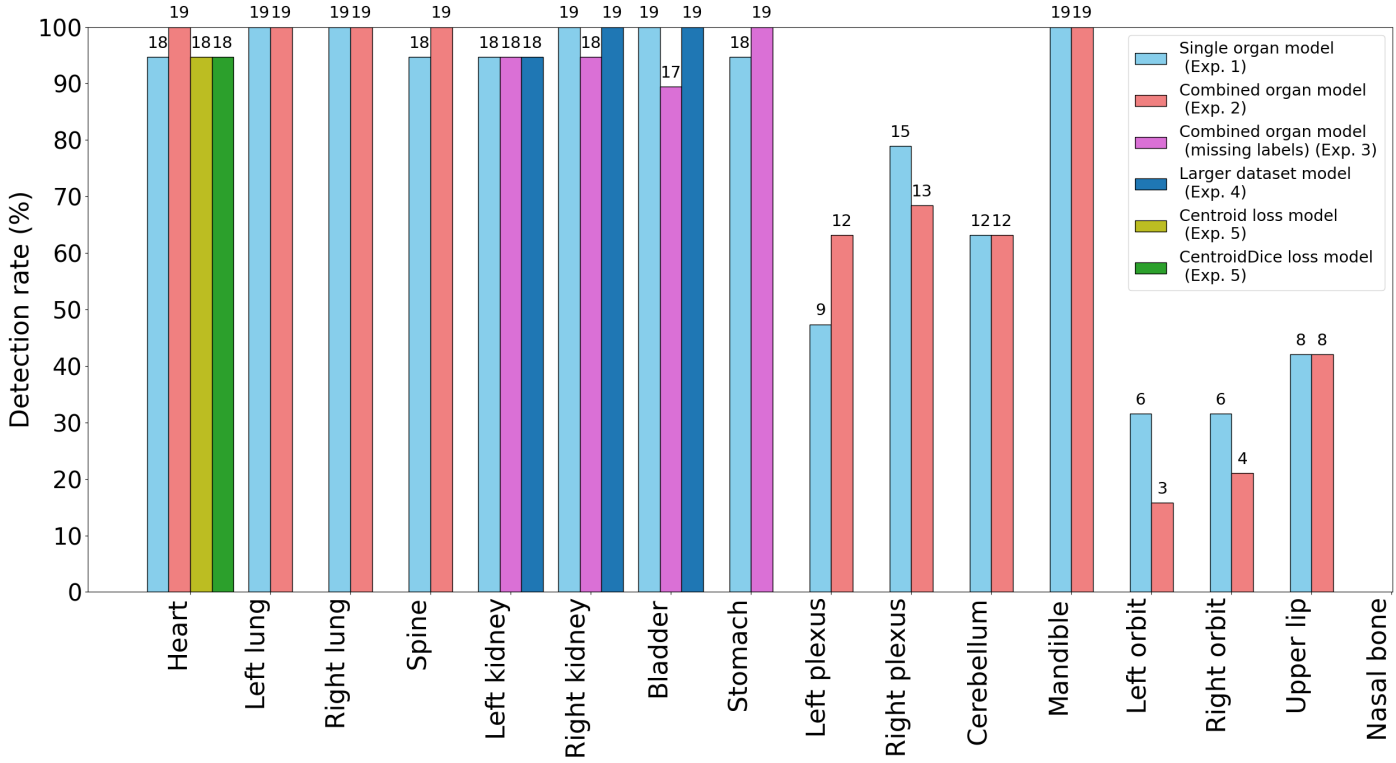


Figure 6: *Detection rate of the model predictions for the organs of interest. The exact number is indicated above the bars, the test set consisted of 19 subjects in total. The bar colors indicate the models: light blue for the single organ models, pink for the combined organ models, purple for the combined organ model with missing labels, dark blue for the models trained with a larger dataset, light green for the model trained with the centroid loss and dark green for the model trained with the centroid Dice loss.*

values of 0.70, 0.73, 0.74 and 0.81 respectively. Table 1 of the Appendix shows an overview of all median DSC, CDM_{norm} and $HD95_{norm}$ values, including the inter quartile ranges (IQR).

4.2. Experiment 2: Combination of labels

In Experiment 2, two separate models were trained with a combination of pseudo labels, one model for the head of the fetus and one model for the body of the fetus. The detection rate for each organ predicted by the combined organ models was similar to the detection rate of the single organ models, see Figure 6. However, for the left plexus the combined organ model predicted more test cases correctly, while for the right plexus the reverse trend was observed. For the left and right orbit the single organ models achieved a higher detection count.

Figure 8 shows a comparison of the CDM_{norm} and $HD95_{norm}$ between the single organ models (Experiment 1) and the combined organ models. The median CDM_{norm} and $HD95_{norm}$ values of the single organ and combined organ models were similar for most organs, except for the choroid plexuses and the orbits. For these organs both metrics had wide distributions and a higher median for the single organ models, especially for the right side. The nasal bone was not detected in any scan of the test set. In Figure 8 significant differences are denoted with an asterisk (*). Notably, significantly better performances for the $HD95_{norm}$ were observed for the single organ models of the left lung, left plexus and the upper lip. An overview of all metrics for the combined organ models is shown in Table 2 of the Appendix.

4.3. Experiment 3: Missing labels

Not all volumes included all organ labels for the bladder, kidneys and stomach. Top cope with this, we trained a model that could handle these missing labels. The

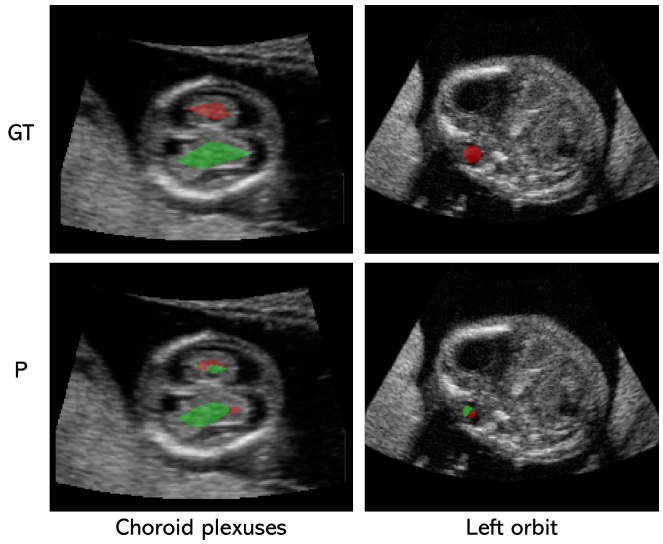


Figure 7: *Examples of test scans for which both left and right labels were predicted on a single side of the organs. The ground truth (GT) and predicted (P) labels are shown for the choroid plexuses and the left orbit.*

model could correctly detect the organs 89 to 100% of the test scans, similar to the single organ models, see Figure 6. Figure 9 shows the DSC for the single organ models and the combined organ model trained with the missing label strategy. The median DSC were $0.72, 0.75$ and 0.72 for the bladder and left and right kidney respectively. These values were similar to the single organ models. However, a significantly worse result was found for the stomach (0.81 vs 0.78). In general the combined organ model trained with the missing label approach did not outperform the single organ models. In Table 2 of the Appendix, an overview of all metrics for the combined organ model trained with missing labels is shown.

4.4. Experiment 4: Dataset size

The models for the bladder and kidneys trained with larger training sets also had a detection rate of 95 to 100%, see Figure 6. The median DSC were $0.72, 0.75$ and 0.74 for the bladder and the left and right kidney, see Figure 10. No significant differences were found between the initial models and those trained with larger training sets. Table 3 of the Appendix shows an overview of the metrics for this experiment.

4.5. Experiment 5: Loss functions

The models for the nasal bone, trained with the $\mathcal{L}_F$ and $\mathcal{L}_F + \mathcal{L}_{SD}$ functions, were not able to detect the nasal bone in any of scans in the test set. For the heart, the models trained with the $\mathcal{L}_{CD}$ and $\mathcal{L}_{CD} + \mathcal{L}_{SD}$ functions, both had a detection rate of 95%, similar to the single organ model (trained with the $\mathcal{L}_{SD} + \mathcal{L}_{CE}$ function). No significant differences were found for the CDM_{norm} and $HD95_{norm}$ between the models trained with different loss functions. Table 4 of the Appendix shows all metrics for these models.

5. Discussion

First trimester anomaly screening is crucial for early detection of congenital disorders. However, it is an operator dependent and time intensive task which needs to be performed by specialists [4]. To overcome these challenges, we trained a deep learning model for automated organ detection in first-trimester 3D ultrasound. Several nnU-Net models were trained during different experiments to determine the configurations that lead to the highest detection rate.

5.1. Experiment outcomes

Across the different experiments for the heart, lungs and spine, we found that the organs were successfully detected in 95% to 100% of the test scans. For the spine, the larger $HD95_{norm}$ for all models could be attributed to the elongated structure of the organ, which might have made the models more susceptible for deviations in the extreme points. For the heart, neither the combined organ model nor the models trained with different loss functions outperformed the single organ model for both the CDM_{norm} and $HD95_{norm}$.

The models for the bladder, kidneys and stomach could detect the organs in 89% to 100% of test scans across Experiment 1, 3 and 4. The high detection rate for Experiment 3 indicates the combined organ model remained robust and achieved similar results to the single organ models, even when some of the organ labels were missing. With regard to the bladder and kidneys, no significant differences were found for the models trained with additional data, as in Experiment 4. While this was the case for the bladder and kidneys, further research should be conducted to determine if

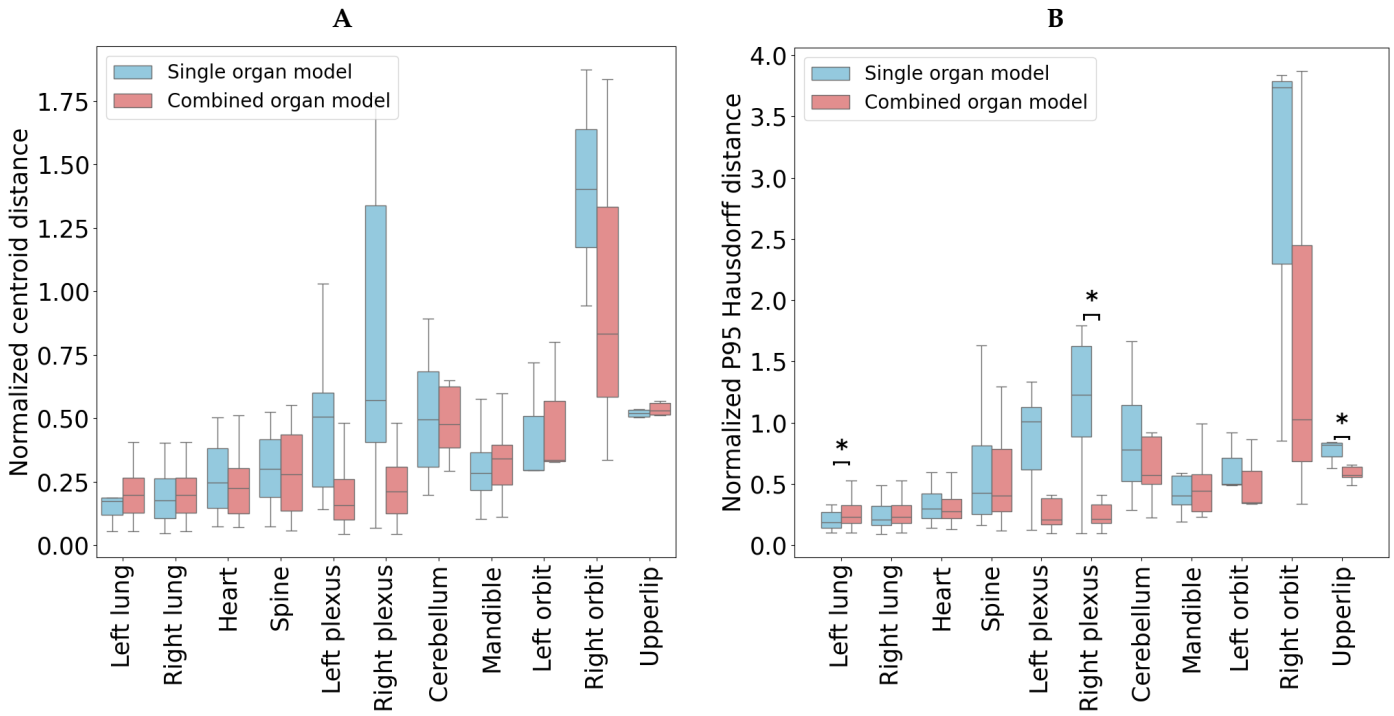


Figure 8: *Comparison of the normalized centroid distances (A) and normalized 95th percentile Hausdorff distances (B) between the single and combined organ models. The results are based on paired test scans that were detected by both models. The blue boxes show the results of the single organ models, the pink boxes show the results for the combined organ models. Significant differences are indicated with an asterisk (*) above the boxes.*

this also holds for different organs in this study.

Both the single and combined organ models detected the choroid plexuses in less than 84% of the test scans. For the orbits, the detection rate was even lower, 32% for the single organ models and less than 26% for the combined organ model. While large differences regarding the CDM_{norm} and $HD95_{norm}$ between the models of the plexuses and orbits were observed in Figure 7, these differences were not significant. A potential cause of this was a limited number of paired samples, which underpowered the Wilcoxon signed-rank test. The large differences in both the CDM_{norm} and $HD95_{norm}$ for the orbits and choroid plexuses was due to the occasional detection of the left and right labels on a single side, as shown in Figure 5. This is probably due to the high similarity in structure and shape of the left and right sides, causing the model to interchange these labels. Interestingly, this happened less often for the combined organ model for the choroid plexuses. This might be because the model had more spatial awareness of what direction the head of the fetus was oriented, taking into account the other organ labels present in the training data.

None of the models were able to detect the nasal bone in the scan of the test set. A possible reason for this is its small size, which creates an in-balance in fore and background. Another reason might be that the nasal bone is still developing in this stage of the pregnancy. During this development, the nasal bone consists of two ossification centers, that later fuse [22]. Our pseudo label for the nasal bone consisted of just one structure, which might have confused the model. Also the accurate placement of the nasal bone annotations is difficult, as the size of the nasal bone ranges between only 1.4mm and 2.1mm during the 11-14 weeks of gestation [23]. Even a minimal shift during the landmark annotation in VR might cause the pseudo label not to coincide with the actual nasal bone, as the mean voxel size was 0.25mm. This can also explain the low detection rate (42% for both models) of the upper lip. This organ is located at the edge of the fetal face, which was sometimes difficult to capture in the ultrasound scan.

Additionally, for all models the variety in scan quality could have played a noticeable role in the performances. Ultrasound artifacts such as movements, shadows and low resolution might have influenced organ visibility. While the nnU-Net framework was designed to account

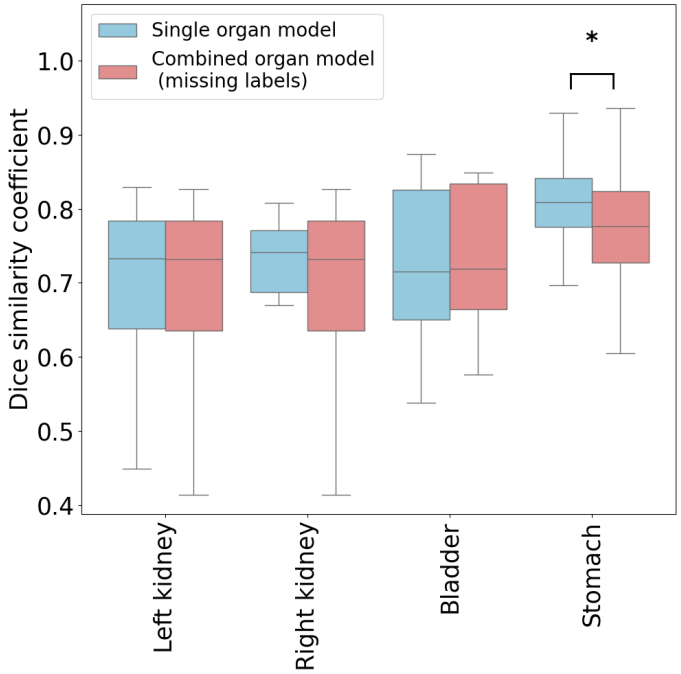


Figure 9: *Comparison of the Dice similarity coefficients for the kidneys, bladder and stomach between the single organ models and the combined organ model with missing labels. The results are based on paired test scans that were detected by both models. Blue boxes show results of the single organ models, pink boxes of the combined organ models with missing labels. Significant differences are indicated with an asterisk (*) above the boxes.*

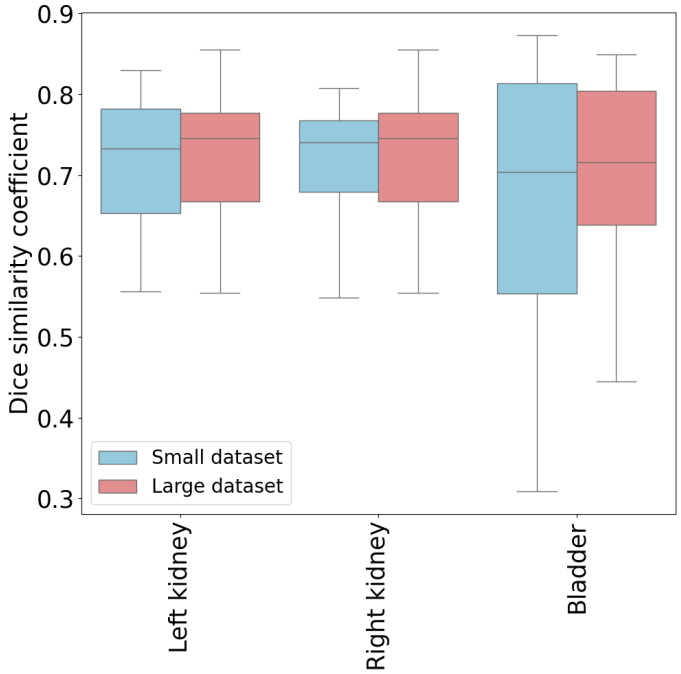


Figure 10: *Comparison of the Dice similarity coefficients for the kidneys and bladder between models trained with a smaller and larger training dataset. The results are based on paired test scans that were detected by both models. Blue boxes show results of the the models trained with a small dataset, pink boxes of the models trained with a larger training set*

for differences in for example organ sizes and intensities, potentially not all variations were captured in the relatively small training set.

5.2. Limitations

In this study we trained models using pseudo labels (except for the bladder, kidneys and stomach), which were constructed using convex hulls, cylinders or spheres. These pseudo labels did not capture all anatomical boundaries. For most organs, training the models with these labels did facilitate detection. However, training with labels that would capture more details of the anatomical boundaries might have enhanced the detection of smaller organs and result in higher performances across all metrics. While full segmentation labels are desirable, their creation is time consuming and complex, due to low resolution and ultrasound artifacts. Despite the absence of full segmentation labels for all organ, we managed to successfully detect the heart, lungs, spine and mandible in at least 95% of the test scans.

We approached a detection problem with a segmentation model, resulting in segmentation mask predictions. This approach introduced challenges for the evaluation, as the predicted masks were evaluated based on the landmarks. We considered an organ detected if the predicted label overlapped with the ground truth label. While this approach yielded promising results, we also aimed to interpret the quality of the detections. We did this by looking at the distances between the centroids of the prediction and ground truth and by looking at the maximum distances using the 95th percentile Hausdorff distance. For organs such as the right choroid plexus and the right orbit, these metrics were quite high, which prompted visual inspection that helped to explain these

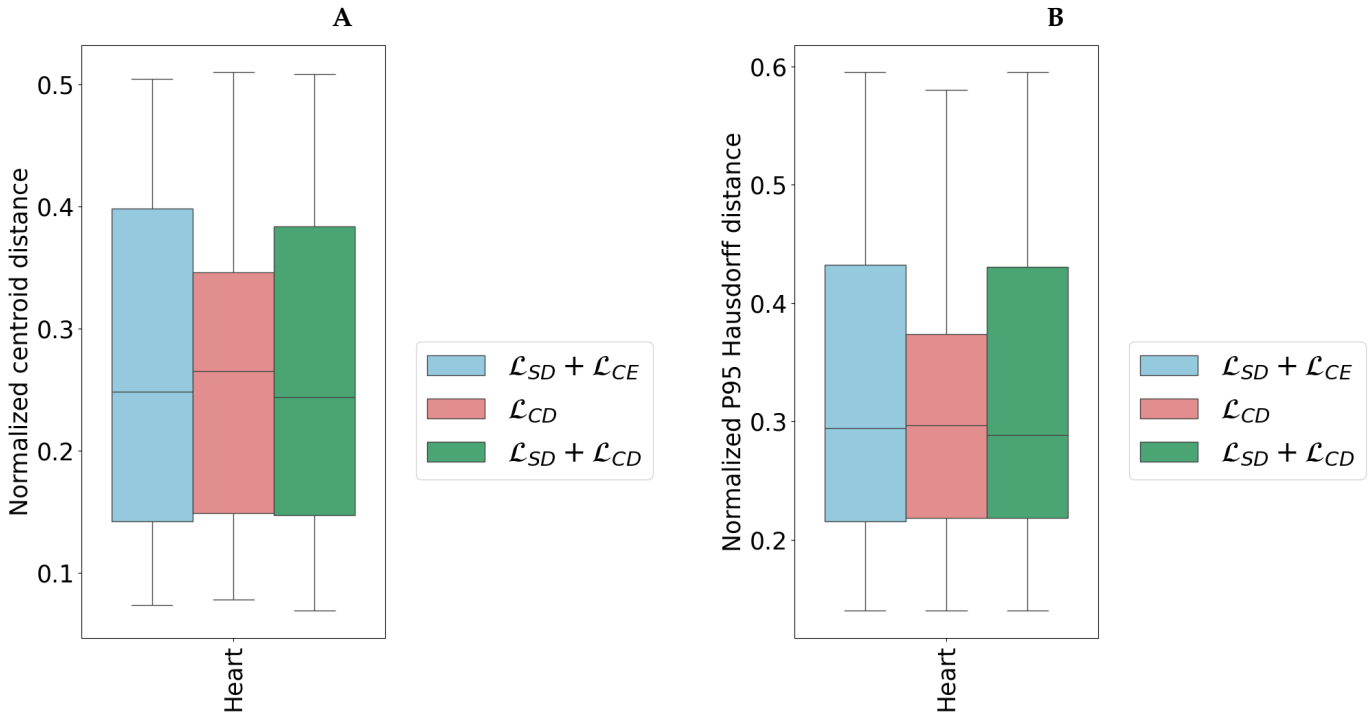


Figure 11: *Comparison of the normalized centroid distance (A) and normalized 95th percentile Hausdorff distance (B) between the ground truth and prediction masks of models for the pseudo label of the heart, trained with different loss functions. The results are based on paired test scans that were detected by both models. The blue boxes show the results of the single model trained with the standard loss function ($\mathcal{L}_{SD} + \mathcal{L}_{CE}$), the pink boxes for the model trained with the $\mathcal{L}_{CD}$ and the green boxes for model trained with the $\mathcal{L}_{SD} + \mathcal{L}_{CE}$.*

results.

Another limitation is that the models were trained and tested using a single dataset that originates from the VR-FETUS study [4]. The 3D ultrasound scans were obtained using the same type of ultrasound machine. To assess generalizibilty, the models should be validated on datasets from different institutes and ultrasound machines. Additionally, the dataset was annotated once by two raters; a fetal medicine specialist and medical student. For a more robust reference standard, the dataset could be annotated by multiple raters, so that the intra-rater variability could be assessed and high annotation consistency could be achieved.

5.3. Future work

In this work, the pseudo labels were modeled through approximate, idealized shapes (spheres, cylinders), which did not actually correspond to the anatomical shapes of the organs. To improve the pseudo labels, the annotated landmarks could be combined with a high quality atlas of first-trimester fetuses. Subsequently, the atlas labels could be registered to the 3D ultrasound volumes, after which the landmarks might be used for refinement. Using this method, pseudo labels that represent a more anatomical shape could be generated. For smaller organs such as the upper lip and the nasal bone, a different approach might be considered. Instead of trying to detect these organs by segmenting them, a detection model that uses bounding boxes as input might improve the detection rate. Additionally, a visual inspection of the detection quality could be performed by a fetal medicine specialist, instead of only looking at the overlap of the ground truth and prediction. A long term future goal would be to use an automatic model that not only indicates if the organ is detected, but also determines if the organ has a normal shape and size. This would facilitate a normality screening instead of an anomaly screening, that would indicate to manually check a structure if an organ of interest deviates from

the normal range.

6. Conclusion

This study is an important step towards automated organ detection for first-trimester anomaly screening in 3D ultrasound. The trained nnU-Net models were able to detect the heart, lungs, spine, stomach, bladder, kidneys and mandible in 89% to 100% of the test scans. The models had difficulty with detection of the choroid plexuses, cerebellum, orbits, upper lip and nasal bone, as these organs were detected in less than 84% of the test scans. Challenges related to the quality of the pseudo labels, the used evaluation metrics and generalizibility of the model require future exploration.

AI statement

For refinement of the writing style of this thesis, ChatGPT-4 was used [24]. It was predominantly used for writing better flowing sentences by using prompts like "Can you make this sentence more flowing", or for improving the scientific sound of sentences, by using prompts like "Can you write this sentence in a more scientific way". All suggestions were carefully reviewed and implemented if suitable.

References

[1] World Health Organization. *Congenital disorders*. 2025. URL: https://www.who.int/health-topics/congenital-anomalies.

[2] Francesca Bardi, Karl Oliver Kagan, and Caterina Maddalena Bilardo. "Firsttrimester screening strategies: A balance between costs, efficiency and diagnostic yield". In: *Prenatal Diagnosis* 43.7 (June 2023), pp. 865–872. ISSN: 0197-3851. DOI: 10.1002/pd.6393. URL: https://obgyn.onlinelibrary.wiley.com/doi/10.1002/pd.6393.

[3] Eline E.R. Lust et al. "Introduction of a nationwide first-trimester anomaly scan in the Dutch national screening program". In: *American journal of obstetrics and gynecology* 232.4 (Apr. 2025), pp. 1–396. ISSN: 10976868. DOI: 10.1016/j.ajog.2024.07.026.

[4] C. S. Pietersma et al. "First trimester anomaly scan using virtual reality (VR FETUS study): Study protocol for a randomized clinical trial". In: *BMC Pregnancy and Childbirth* 20.1 (Sept. 2020). ISSN: 14712393. DOI: 10.1186/s12884-020-03180-8.

[5] C. M. Bilardo et al. "ISUOG Practice Guidelines (updated): performance of 1114week ultrasound scan". In: *Ultrasound in Obstetrics & Gynecology* 61.1 (Jan. 2023), pp. 127–143. ISSN: 0960-7692. DOI: 10.1002/uog.26106. URL: https://obgyn.onlinelibrary.wiley.com/doi/10.1002/uog.26106.

[6] G D Michailidis, P Papageorgiou, and D L Economides. "Assessment of fetal anatomy in the first trimester using two- and three-dimensional ultrasound". In: *The British Journal of Radiology* 75.891 (Mar. 2002), pp. 215–219. ISSN: 0007-1285. DOI: 10.1259/bjr.75.891.750215. URL: https://academic.oup.com/bjr/article/75/891/215-219/7443179.

[7] Eberhard Merz and C. Welter. "2D and 3D ultrasound in the evaluation of normal and abnormal fetal anatomy in the second and third trimesters in a level III center". In: *Ultraschall in der Medizin* 26.1 (Feb. 2005), pp. 9–16. ISSN: 01724614. DOI: 10.1055/s-2004-813947.

[8] Leonie Baken et al. "First-Trimester Detection of Surface Abnormalities: A Comparison of 2- and 3-Dimensional Ultrasound and 3-Dimensional Virtual Reality Ultrasound". In: *Reproductive Sciences* 21.8 (Aug. 2014), pp. 993–999. ISSN: 1933-7191. DOI: 10.1177/1933719113519172.

[9] Caroline Raynaud et al. "Multi-organ Detection in 3D Fetal Ultrasound with Machine Learning". In: *Lecture notes in computer science*. Sept. 2017, pp. 62–72. DOI: 10.1007/978-3-319-67561-9{_}7.

[10] Ruobing Huang, Weidi Xie, and J. Alison Noble. "VP-Nets: Efficient automatic localization of key brain structures in 3D fetal neurosonography". In: *Medical Image Analysis* 47 (July 2018), pp. 127–139. ISSN: 13618423. DOI: 10.1016/j.media.2018.04.004.

[11] Guang Quan Zhou et al. "Learn Fine-Grained Adaptive Loss for Multiple Anatomical Landmark Detection in Medical Images". In: *IEEE Journal of Biomedical and Health Informatics* 25.10 (Oct. 2021), pp. 3854–3864. ISSN: 21682208. DOI: 10.1109/JBHI.2021.3080703.

[12] Genta Ishikawa et al. "Detecting a Fetus in Ultrasound Images using Grad CAM and Locating the Fetus in the Uterus". In: *International Conference on Pattern Recognition Applications and Methods*. Vol. 1. Science and Technology Publications, Lda, 2019, pp. 181–189. ISBN: 9789897583513. DOI: 10.5220/0007385001810189.

[13] T. R. Nelson et al. "Sources and impact of artifacts on clinical three-dimensional ultrasound imaging". In: *Ultrasound in Obstetrics and Gynecology* 16.4 (2000), pp. 374–383. ISSN: 09607692. DOI: 10.1046/j.1469-0705.2000.00180.x.

[14] Zhong-Qiu Zhao et al. "Object Detection With Deep Learning: A Review". In: *IEEE Transactions on Neural Networks and Learning Systems* 30.11 (Nov. 2019), pp. 3212–3232. ISSN: 2162-237X. DOI: 10.1109/TNNLS.2018.2876865. URL: https://ieeexplore.ieee.org/document/8627998/.

[15] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool". In: *BMC Medical Imaging* 15.1 (Aug. 2015). ISSN: 14712342. DOI: 10.1186/s12880-015-0068-x.

[16] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (Feb. 2021), pp. 203–211. ISSN: 15487105. DOI: 10.1038/s41592-020-01008-z.

[17] Wietske A.P. Bastiaansen et al. "Automatic Human Embryo Volume Measurement in First Trimester Ultrasound From the Rotterdam Periconception Cohort: Quantitative and Qualitative Evaluation of Artificial Intelligence". In: *Journal of Medical Internet Research* 27 (2025). ISSN: 14388871. DOI: 10.2196/60887.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Lecture Notes in Computer Science* (May 2015), pp. 234–241. DOI: 10.1007/978-3-319-24574-4{_}28. URL: http://arxiv.org/abs/1505.04597.

[19] Koning Anton H.J. et al. "V-Scope: Design and Implementation of an Immersive and Desktop Virtual Reality Volume Visualization System". In: *Studies in Health Technology and Informatics*. Vol. 142. 2009. DOI: 10.3233/978-1-58603-964-6-136.

[20] Juan Terven et al. "A comprehensive survey of loss functions and metrics in deep learning". In: *Artificial Intelligence Review* 58.7 (July 2025). ISSN: 15737462. DOI: 10.1007/s10462-025-11198-7.

[21] Jason Wong et al. "Centroid-based Distance Loss Function for Lamina Segmentation in 3D Ultrasound Spine Volumes". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, Nov. 2021, pp. 1723–1726. ISBN: 978-1-7281-1179-7. DOI: 10.1109/EMBC46164.2021.9631034.

[22] In-Sang Kim et al. "Analysis of the Development of the Nasal Septum according to Age and Gender Using MRI". In: *Clinical and Experimental Otorhinolaryngology* 1.1 (2008), p. 29. ISSN: 1976-8710. DOI: 10.3342/ceo.2008.1.1.29.

[23] Chitkasaem Suwanrath et al. "Reliability of fetal nasal bone length measurement at 11-14 weeks of gestation". In: *BMC Pregnancy and Childbirth* 13 (Jan. 2013). ISSN: 14712393. DOI: 10.1186/1471-2393-13-7.

[24] OpenAI. *ChatGPT-4*. 2023. URL: https://chat.openai.com/.

Appendix

Table 1: *Evaluation metrics for the single organ models. The detection rate (DR), median Dice similarity coefficient (DSC), median normalized centroid distance (CDM_{norm}) and median normalized 95th percentile ($HD95_{norm}$) are shown with the interquartile range (IQR) between brackets.*

Organ	DR (%)	DSC (IQR)	CDM_{norm} (IQR)	$HD95_{norm}$ (IQR)
Bladder	100	0.70 (0.67-0.72)	0.18 (0.16-0.19)	0.28 (0.18-0.47)
Kidney (L)	95	0.73 (0.72-0.76)	0.18 (0.16-0.20)	0.17 (0.16-0.20)
Kidney (R)	100	0.74 (0.74-0.74)	0.16 (0.16-0.16)	0.18 (0.18-0.18)
Stomach	95	0.81 (0.80-0.82)	0.06 (0.05-0.08)	0.15 (0.15-0.16)
Cerebellum	63	0.16 (0.10-0.22)	0.50 (0.40-0.59)	0.86 (0.78-1.03)
Heart	95	0.64 (0.59-0.66)	0.25 (0.19-0.30)	0.29 (0.26-0.33)
Lung (L)	100	0.73 (0.73-0.75)	0.18 (0.15-0.18)	0.18 (0.18-0.18)
Lung (R)	100	0.67 (0.67-0.72)	0.18 (0.17-0.21)	0.21 (0.19-0.21)
Mandible	100	0.37 (0.36-0.37)	0.28 (0.27-0.29)	0.40 (0.38-0.47)
Orbit (L)	32	0.33 (0.28-0.37)	0.33 (0.30-0.51)	0.50 (0.50-0.71)
Orbit (R)	32	0.32 (0.23-0.40)	1.17 (0.95-1.41)	2.30 (0.85-3.74)
Plexus (L)	47	0.33 (0.14-0.51)	0.51 (0.44-0.54)	1.01 (0.71-1.11)
Plexus (R)	79	0.46 (0.43-0.47)	0.68 (0.62-0.97)	1.25 (1.21-1.31)
Spine	95	0.44 (0.41-0.48)	0.30 (0.27-0.33)	0.43 (0.36-0.52)
Upperlip	42	0.14 (0.09-0.17)	0.53 (0.52-0.58)	0.83 (0.82-0.84)

Table 2: *Evaluation metrics for the combined organ models. The detection rate (DR), median Dice similarity coefficient (DSC), median normalized centroid distance (CDM_{norm}) and median normalized 95th percentile ($HD95_{norm}$) are shown with the interquartile range (IQR) between brackets.*

Organ	DR (%)	DSC (IQR)	CDM_{norm} (IQR)	$HD95_{norm}$ (IQR)
Bladder	89	0.72 (0.72-0.73)	0.17 (0.15-0.17)	0.26 (0.23-0.28)
Kidney (L)	95	0.72 (0.70-0.73)	0.19 (0.17-0.22)	0.19 (0.18-0.20)
Kidney (R)	95	0.75 (0.73-0.77)	0.15 (0.14-0.17)	0.17 (0.16-0.18)
Stomach	100	0.78 (0.77-0.78)	0.09 (0.08-0.10)	0.17 (0.16-0.18)
Cerebellum	63	0.24 (0.22-0.24)	0.48 (0.41-0.55)	0.56 (0.54-0.57)
Heart	100	0.71 (0.63-0.71)	0.23 (0.19-0.23)	0.23 (0.23-0.24)
Lung (L)	100	0.67 (0.66-0.73)	0.20 (0.19-0.20)	0.20 (0.20-0.24)
Lung (R)	100	0.64 (0.63-0.64)	0.23 (0.22-0.25)	0.28 (0.27-0.28)
Mandible	100	0.42 (0.36-0.45)	0.34 (0.27-0.35)	0.44 (0.38-0.45)
Orbit (L)	16	0.19 (0.06-0.52)	0.54 (0.33-0.80)	0.79 (0.35-0.86)
Orbit (R)	21	0.19 (0.07-0.31)	0.76 (0.69-1.09)	0.95 (0.87-2.11)
Plexus (L)	63	0.64 (0.61-0.67)	0.23 (0.21-0.25)	0.21 (0.21-0.24)
Plexus (R)	68	0.62 (0.59-0.63)	0.28 (0.26-0.43)	0.30 (0.22-0.41)
Spine	100	0.40 (0.40-0.45)	0.31 (0.25-0.32)	0.41 (0.40-0.59)
Upper lip	42	0.21 (0.20-0.25)	0.51 (0.49-0.53)	0.56 (0.53-0.57)

Table 3: *Evaluation metrics for models for the bladder and kidneys trained with different dataset sizes. The detection rate (DR) and Dice similarity coefficient (DSC) with the interquartile range (IQR) between brackets are shown.*

Organ	Model	DR (%)	DSC (IQR)
Bladder	Small trainingset	100	0.70 (0.67-0.72)
	Large traininset	100	0.72 (0.71-0.72)
Kidney (L)	Small trainingset	95	0.73 (0.72-0.76)
	Large traininset	95	0.75 (0.74-0.76)
Kidney (R)	Small trainingset	100	0.74 (0.74-0.74)
	Large traininset	100	0.74 (0.74-0.75)

Table 4: *Metrics for models for the heart trained with different loss functions. The detection rate (DR), normalized centroid distance (CDM_{norm}) and 95th percentile normalized Hausdorff distance ($HD95_{norm}$) are shown. The interquartile range (IQR) for the last two metrics is shown between brackets.*

Loss function	DR (%)	CDM_{norm} (IQR)	$HD95_{norm}$ (IQR)
$\mathcal{L}_{CE} + \mathcal{L}_{SD}$	95	0.31 (0.23-0.43)	0.16 (0.14-0.22)
$\mathcal{L}_{CD}$	95	0.27 (0.21-0.31)	0.30 (0.26-0.34)
$\mathcal{L}_{CD} + \mathcal{L}_{SD}$	95	0.24 (0.19-0.30)	0.29 (0.26-0.33)