Published in final edited form as:

Med Image Anal. 2023 April ; 85: 102731. doi:10.1016/j.media.2022.102731.

Learning to segment fetal brain tissue from noisy annotations

Davood Karimi[a,*], **Caitlin K. Rollins**[b], **Clemente Velasco-Annis**[a], **Abdelhakim Ouaalam**[a], **Ali Gholipour**[a]

[a]Department of Radiology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

[b]Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

this paper is about automatically segment different brain tissue even in with noisy image in dataset

Abstract

Automatic fetal brain tissue segmentation can enhance the quantitative assessment of brain development at this critical stage. Deep learning methods represent the state of the art in medical image segmentation and have also achieved impressive results in brain segmentation. However, effective training of a deep learning model to perform this task requires a large number of training images to represent the rapid development of the transient fetal brain structures. On the other hand, manual multi-label segmentation of a large number of 3D images is prohibitive. To address this challenge, we segmented 272 training images, covering 19–39 gestational weeks, using an automatic multi-atlas segmentation strategy based on deformable registration and probabilistic atlas fusion, and manually corrected large errors in those segmentations. Since this process generated a large training dataset with noisy segmentations, we developed a novel label smoothing procedure and a loss function to train a deep learning model with smoothed noisy segmentations. Our proposed methods properly account for the uncertainty in tissue boundaries. We evaluated our method on 23 manually-segmented test images of a separate set of fetuses. Results show that our method achieves an average Dice similarity coefficient of 0.893 and 0.916 for the transient structures of younger and older fetuses, respectively. Our method generated results that were significantly more accurate than several state-of-the-art methods including nnU-Net that achieved the closest results to our method. Our trained model can serve as a valuable tool to enhance the accuracy and reproducibility of fetal brain analysis in MRI.

Keywords

fetal brain; tissue segmentation; noisy labels; deep learning

*Corresponding author: Tel.: +1-617-208-9736; fax: +1-617-730-0635.

The authors do not have any conflicts of interest (financial or otherwise) to dosclose.

The **Dice Similarity Coefficient (DSC)** — also called the **Sørensen–Dice coefficient** — is a statistical measure used to calculate how similar two sets are. It is widely used in **image segmentation**, **clustering**, and **natural language processing** to evaluate overlap between predicted results and ground truth.

✅ Definition

For two sets **A** and **B**, the Dice coefficient is:

$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|}$$

✅ Intuition

- Measures **overlap** between two samples.
- **0** means *no overlap*.
- **1** means *perfect overlap*.

🧠 In Image Segmentation

A common application is comparing how well an algorithm segments an organ or object compared to manual (ground truth) segmentation.

If **X** = predicted pixels
and **Y** = ground-truth pixels:

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

Where

- **TP** = True Positives
- **FP** = False Positives
- **FN** = False Negatives

$$\text{DSC} = \frac{2 \cdot \text{IoU}}{1 + \text{IoU}}$$

- To do quantitative analysis (volumes, shapes, growth curves), you need to **segment** the brain into many tissue types: cortical plate, subplate, ventricles, white matter, deep nuclei, cerebellum, etc.
- Doing this **manually in 3D** for many fetuses is:
  - Extremely time-consuming (days per scan)
  - Hard, because fetal MRI has motion artifacts, partial volume effects, and rapidly changing anatomy with gestational age.

Main difficulty

- Deep learning needs **lots of labeled images**, but:
  - Getting *perfect* 3D multi-label segmentations is basically impossible at large scale.
  - Even "good" labels still have **uncertain boundaries** (especially for thin or poorly contrasted structures).
- So they accept that their labels will be **noisy** and **uncertain near tissue boundaries**, and try to **explicitly model and use that uncertainty** during training instead of ignoring it.

Core idea

1. Use an **automatic atlas-based method** to segment a large training set, then **manually correct only big mistakes** → results are *noisy* but cheap to obtain.
2. Ask experts to specify, for each tissue type, **how uncertain its boundary is** (in voxels).
3. Use this information to:
   - Create **spatially varying "soft" labels** near boundaries (label smoothing).
   - Design a **loss function that treats certain and uncertain voxels differently**, using a learned **label transition matrix** that describes how labels tend to get confused.
4. Train a 3D U-Net (via nnU-Net framework) on 272 noisy scans using this strategy.
5. Evaluate on 22 fetuses with **carefully manually segmented ground truth** and compare to many alternative methods.

They show that this approach **beats nnU-Net and other strong baselines**, and even outperforms a classical multi-atlas segmentation method. nihms-1861994

→ they used soft labeling for this, instead of one hot vector for labeling (Realistic data) they provide vector with probability of belonging to classes

these are regions in brain

Because of brain development:

- **Younger fetuses (<32 weeks)**: SP and IZ are separate labels.
- **Older fetuses (≥32 weeks)**: SP + IZ are merged into WM.
- So they train **two separate models**:
  - Young: **33 tissue labels**
  - Old: **31 tissue labels** (plus background in both cases)

How the *test* labels are created (high-quality "clean" labels)
For the 22 test scans:

1. Use a **spatiotemporal fetal brain atlas** (covering 19–39 weeks) as prior segmentations.
2. For each test fetus:
   - Register atlas images within ±1 week of that fetus's GA using **diffeomorphic deformable registration.**
   - Propagate the atlas labels to the subject.
3. Combine multiple warped labels using **probabilistic STAPLE** (a label-fusion method). nihms-1861994
4. Then **expert annotators manually refine *all* labels thoroughly**, in several rounds, until they are "clean".

This gives them **very accurate ground truth** for evaluation, but takes **4–10 days per scan**.

How the *training* labels are created (noisy labels)
For the 272 training scans, they do a similar pipeline:

1. Same multi-atlas segmentation (registration + STAPLE).
2. Annotators **correct only major errors**, not fine details → ~2 hours per scan.

So the training labels:

- Are pretty good overall.
- But still have **boundary noise**, especially for structures that are hard to see or thin.

Crucially, annotators also provide for each tissue:

- A single number: **boundary uncertainty**, measured in voxels (0–4), e.g.:
  - Lateral ventricle: uncertainty = 0 voxels (sharp boundary).
  - Caudate nuclei: uncertainty = 2 voxels (blurry / uncertain edge).

This is the key input to their label smoothing method. nihms-1861994

1. Introduction

1.1. Background and motivation

Fetal magnetic resonance imaging (MRI) has emerged as an important and viable tool for assessing the development of brain in utero. It has enabled assessment of normal and abnormal brain growth trajectories in utero (Corbett-Detig et al., 2011). Moreover, fetal MRI may offer more accurate assessment and quantification of fetal brain development and degeneration when ultrasound images are inadequate (Hosny and Elghawabi, 2010; Weisstanner et al., 2015). Faster image acquisition methods (Yamashita et al., 1997) and superior super-resolution algorithms (Ebner et al., 2020; Kainz et al., 2015; Kuklisova-Murgasova et al., 2012; Gholipour et al., 2010) can now reconstruct high-quality 3D fetal brain images from stacks of 2D slices. These technical advancements have significantly improved the quality of fetal brain MRI. As a result, a growing number of works have successfully used fetal MRI to study various congenital brain disorders (Egaña-Ugrinovic et al., 2013; Mlczoch et al., 2013). As the use of fetal MRI in clinical and research studies grows, quantitative image analysis methods become an urgent requirement. Automatic analysis methods can increase the speed, accuracy and reproducibility of quantification of fetal brain development. Accurate segmentation of the fetal brain into relevant tissue compartments is especially critical because many congenital brain disorders manifest themselves as changes in the size or shape of these tissues. Whereas manual segmentation is time-consuming and prone to high intra/inter-observer variability (Gousias et al., 2012), automatic segmentation promises high speed and reproducibility. As a result, in recent years there have been multiple efforts to segment different tissue compartments in the fetal brain.

review Paper about brain segmentation

Makropoulos et al. (2018) have reviewed automatic fetal and neonatal brain segmentation techniques in MRI. Here, we focus primarily on deep learning (DL) methods. Deep learning-based segmentation of adult brain into different tissue compartments has been successfully attempted by several studies in recent years (Dolz et al., 2018; Sun et al., 2019). Overall, they show that DL methods are capable of accurately segmenting brain into relevant tissue compartments. Comparatively, much fewer studies have targeted the fetal brain tissue segmentation. For fetal cortical gray matter segmentation, one study proposed obtaining cheap annotations using an automatic segmentation method originally designed for neonatal brains (Fetit et al., 2020). They used a human-in-the-loop method to refined those segmentations on selected 2D slices. A fully convolutional network (FCN) was trained using these annotations. Their method achieved an average Dice Similarity Score (DSC) of 0.76. Segmentation of cortical gray matter has been addressed by several other works. One work used a deep attentive FCN and reported a mean DSC of 0.87 (Dou et al., 2020), whereas another study proposed integrating a topological constraint into the training loss function and achieved a mean DSC of 0.70 (Dumast et al., 2021). Another study used the nnU-Net framework to segment the white matter, ventricles, and cerebellum in fetal brain MRI, achieving DSC values in the range 0.78–0.94 (Fidon et al., 2021). To reduce the impacts of motion artifacts and partial volume effects, Li et al. (2021) proposed a unified deep learning framework to jointly estimate a conditional atlas and predict a segmentation. The rationale for this approach is that the prior knowledge provided by the atlas can guide the segmentation where image quality is low. The idea of leveraging atlases to improve

human in loop

* very important Point as boundaries in fetal MRIs are low contrast

Human-in-the-loop (HITL) refers to a workflow where **an automated system does the initial work**, and then **a human expert reviews, corrects, or guides the system**, creating a loop between machine and human until the result is sufficiently good.

1. What usually happens in human-in-the-loop segmentation?

A typical HITL workflow in medical imaging looks like this:

1. **Step 1 – Automatic segmentation**
   A preliminary segmentation is generated by an algorithm (e.g., atlas-based, deep learning, or a neonatal segmentation method as in the cited study).
2. **Step 2 – Human correction**
   A radiologist or trained annotator:
   - Checks specific slices
   - Edits mistakes (e.g., missing regions, mislabeled boundaries)
   - Deletes false positives, fills holes, fixes anatomical aliasing
3. **Step 3 – Model updates or acceptance**
   - The corrections can be used to **improve the model** (i.e., retraining)
   - Or simply accepted to produce a higher-quality final segmentation
4. **Step 4 – Loop continues**
   If the algorithm is retrained on refined labels, the next automatic segmentation is slightly better → requires fewer corrections next time.

This is why it's called a *loop*.

3. What exactly happened in the cited study?

The paper refers to another work (Fetit et al., 2020), where:

- They used a neonatal brain segmentation tool to produce rough fetal segmentations.
- These segmentations were often wrong because neonatal models don't generalize perfectly to fetuses.
- Instead of manually segmenting entire 3D brains:
  - Annotators corrected *only selected 2D slices* that were most problematic.
  - Their corrections were used to refine or retrain the model.

Thus, HITL allowed them to cheaply obtain better labels for training a fetal cortical plate segmentation network.

- They use an **atlas-based automatic segmentation** to generate initial labels.
- Annotators only **fix major errors**, not all details.
  (This is their own version of HITL.)
- They then use these *inexpensive but noisy labels* to train a deep learning model robustly.

deep learning-based segmentation has been explored in several other works (Oguz et al., 2018; Diniz et al., 2020; Karimi et al., 2018; Zeng et al., 2018; Karimi et al., 2019). A succession of two FCNs was proposed by Khalili et al. (2019), the first to extract the intracranial volume and the second to segment the brain tissue into seven compartments. This method achieved a mean DSC of 0.88. For segmenting the fetal brain into seven tissue compartments, another study used a single 2D UNet and achieved a mean DSC of 0.86 (Payette et al., 2020). Payette et al. (2021) compared a multi-atlas segmentation method with several DL methods for segmentation of fetal brain into seven tissue types and found that overall DL methods can achieve more accurate results.

1.2. Segmentation with noisy labels

Deep learning segmentation models, which represent the state of the art, require large accurately-labeled training datasets. Such datasets are especially difficult to come by in fetal MRI because the image quality is low and accurate multi-label segmentation of 3D images is very time-consuming. Despite recent progress in super-resolution reconstruction methods, 3D fetal MR images can suffer from residual motion and partial volume effects, making accurate delineation of tissue boundaries challenging and uncertain. Moreover, the fetal brain undergoes rapid and significant changes during the second and the third trimesters. Therefore, to develop an accurate DL model, training data should include a sufficiently large number of subjects at different gestational ages (GA) in order to fully capture the variability in the transient fetal brain structures.

Because detailed manual segmentation of a large number of 3D fetal brain images is impossible or prohibitive, an alternative strategy would be to use less accurate annotations. Scenarios with weak, partial, or noisy labels are very common in medical image analysis. Hence, training of DL models with imperfect labels has been the subject of intense research in recent years (Cheplygina et al., 2019; Tajbakhsh et al., 2019; Rajchl et al., 2016). Such labels can often be obtained at low cost using automatic or semi-automatic methods. Song et al. (2020) have reviewed the state of the art methods for handling the label noise in DL. Karimi et al. (2020) present a survey that is more focused on medical image analysis applications, where the authors have identified six classes of methods for training DL models under strong label noise. Below, we describe two of the techniques that are more relevant to this work.

Loss function.—There have been many efforts to devise loss functions that are tolerant to label noise (Zhang and Sabuncu, 2018; Rusiecki, 2019). These loss functions typically tend to down-weight the penalty on data samples that incur very high loss values, under the assumption that those data samples are likely to have wrong labels. Another group of loss functions and training procedures are based on estimating and incorporating a label transition matrix (Patrini et al., 2017; Sukhbaatar et al., 2014). Label transition matrix $T \in \mathbb{R}^{L \times L}$ where L is the number of labels, is meant to describe how the correct labels are flipped into incorrect labels. If we denote the clean and noisy class probability vectors with, respectively, $p_c \in \mathbb{R}^L$ and $p_n \in \mathbb{R}^L$, then we have $p_n = T p_c$. Hence, $T_{i,j}$ is the probability that the correct label j is flipped to label i. Different approaches to estimating T and using it in

training DL classification models have been proposed in prior works (Thekumparampil et al., 2018; Bekker and Goldberger, 2016).

Label smoothing.—Label smoothing has been extensively used in image classification (Pereyra et al., 2017) as well as in natural language processing applications (Chorowski and Jaitly, 2016). However, very few studies have used label smoothing for segmentation applications. In fact, the standard label smoothing approach is unlikely to be suitable for segmentation applications. This is because, unlike classification where the whole image is represented with one probability vector, in segmentation a probability vector belongs to a single pixel/voxel and there are strong spatial correlations between the labels of nearby voxels. Standard label smoothing ignores those spatial correlations and essentially assumes that the probability that label k is flipped to label $l \neq k$ is the same for all l, which is an unrealistic assumption. One study suggested smoothing the object boundaries in training data in order to improve the uncertainty calibration of the trained model (Islam and Glocker, 2021). However, although they used the term “spatially-varying” to describe their method, they applied a fixed operation to all voxels. Another study proposed a label smoothing approach to improve the model uncertainty calibration for scene segmentation (Liu et al., 2021). In the context of image colorization, one study used label smoothing to achieve more accurate scene segmentation (Nguyen-Quynh et al., 2020). However, none of these studies have properly addressed the spatially-varying nature of boundary uncertainty in semantic segmentation.

Another challenge in fetal brain tissue segmentation is that it involves a large number of compartments that vary significantly in size. In this work, we aim to segment the fetal brain into more than 30 classes, where the volume of the smallest class is typically 10^4 times smaller than the volume of the largest class. This can present a significant challenge for some of the loss functions that are commonly used to train DL segmentation models.

The goal of this work is to develop methods for accurate fetal brain tissue segmentation in MRI. Most prior works have segmented only a single tissue (e.g., Dou et al. (2020); Fetit et al. (2020)) or have divided the fetal brain into a small number of tissue compartments (e.g., Payette et al. (2020); Fidon et al. (2021)). In this work, we consider more than 30 relevant and important tissue compartments. To capture the rapid brain growth in utero and the complex developmental trajectories of these tissues, we use a training dataset of 272 images covering the gestational age between 19 and 39 weeks. Instead of manually annotating this dataset, which would have been prohibitive, we use a combination of automatic atlas-based segmentation and manual correction of gross errors. Using the expert-estimated boundary uncertainty for different tissues, we develop novel methods for label smoothing and for training a DL model with the smoothed labels. We evaluate our trained model and compare it with several alternative methods on a set of manually-segmented test images. We show that our method achieves high segmentation accuracy and outperforms several state of the art methods.

2. Materials and methods

2.1. Data and annotation procedures

Data from 294 fetuses with GA between 19.6 and 38.9 weeks (mean 30.6; standard deviation 5.3) were used in this study. These data were collected in studies approved by the institutional review board committee. Written informed consent was obtained from pregnant women volunteers who participated in research MRI scans for fetal MRI. All images were collected with 3-Tesla Siemens Skyra, Trio, or Prisma scanners using 18 or 30-channel body matrix coils via repeated T2-weighted half-Fourier acquisition single shot fast spin echo (T2wSSFSE) scans in the orthogonal planes of the fetal brain. The slice thickness was 2mm with no inter-slice gap, in-plane resolution was between 0.9mm and 1.1mm, and acquisition matrix size was 256 × 204, 256 × 256, or 320 × 320. Volumetric images were reconstructed using an iterative slice-to-volume reconstruction algorithm (Kainz et al., 2015), brain extracted and registered to a standard atlas space in a procedure described in (Gholipour et al., 2017). The resulting 3D images had isotropic voxels of size 0.8mm. We selected 22 of these fetuses and set them aside as “test subjects” for final evaluation. The test subjects had GA between 23.3 and 38 weeks (mean 32.9; standard deviation 4.14). We used the remaining, completely independent, 272 subjects as “training subjects” to develop/train our methods and also to train the competing techniques.

We manually segmented the test images in detail. To speed up the process, we first generated automatic segmentation for each subject with a multi-atlas segmentation method using a publicly available atlas (Gholipour et al., 2017). This is a four-dimensional (i.e., spatio-temporal) atlas that covers the GA range between 19 and 39 weeks at one-week intervals. For each test fetus, we registered atlases that were within one week GA of the fetus using a diffeomorphic deformable registration algorithm. We then used the probabilistic Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm (Akhondi-Asl and Warfield, 2013) to fuse the segmentations. Then, experienced annotators carefully refined all labels in several rounds until the segmentations were consistent and free of any non-trivial errors. This was a laborious effort, which required 4–10 days of work for each scan. We used these manual segmentations as “ground truth” to test our method and competing techniques.

We then segmented the 272 training scans using a similar two-step approach, but with one major difference. Specifically, in the manual refinement step the annotators only corrected major errors, which on average required approximately two hours of work for each scan that had major errors. This was done because manually segmenting all 272 images with the same level of detail as done for the test images would have been impossible given the annotators’ time. Furthermore, in order to account for the potential errors and uncertainties in these segmentations, we asked the annotators to specify the degree of uncertainty in the boundary of each tissue type. Given that all images had the same spatial resolution, this uncertainty was expressed in terms of the number of voxels. The boundary uncertainty specified by the annotators varied considerably for different tissues. For example, for lateral ventricle the boundary uncertainty was 0 voxels, meaning that the boundary was generally unambiguous, while for the caudate nuclei it was 2 voxels.

Labels considered in this study included the following: hippocampus (HP)†, amygdala (AM)†, caudate nuclei (CD)†, lentiform nuclei (LN)†, thalami (TH)†, corpus callosum (CC), lateral ventricles (LV)†, brainstem (ST), cerebellum (CR)†, subthalamic nuclei (SN)†, hippocampal commissure (HC), fornix (FN), cortical plate (CP)†, subplate zone (SP)†, intermediate zone (IZ)†, ventricular zone (VZ)†, white matter (WM)†, internal capsule (IC)†, CSF, and ganglionic eminence (GE)†. A † next to a label in this list indicates that separate components in the left and the right brain hemispheres were considered for that tissue type. In the rest of this paper, we use the acronyms defined above to refer to these tissues and structures. Following the construction of the atlas (Gholipour et al., 2017), there is one age-dependent difference in tissue labels. Specifically, fetuses that are younger than 32 weeks GA have separate SP and IZ labels, whereas for fetuses that are 32 weeks GA and older these two tissue types are merged as a single label: WM. As a result, younger fetuses have 33 tissue labels, whereas older fetuses have 31 labels (in addition to the background label). Therefore, for our method and also for all competing methods, we trained two separate models for the two fetal age groups.

2.2. Development of a DL-based segmentation method

2.2.1. Label smoothing.—In order to account for the inherent and unavoidable uncertainty in the tissue boundaries, we propose a spatially-varying label smoothing method. Our label smoothing method is presented in Algorithm 1. It is based on the fact that label uncertainty is limited to tissue boundaries and depends on the tissues that meet at the boundary. Specifically, our label smoothing strategy is based on the observation that, as confirmed by our annotators, the boundary uncertainty is dictated by the *more certain* tissue. For example, consider a boundary where one of the adjoining tissues has an uncertainty of two voxels (less certain) but the other has an uncertainty of zero voxels (more certain). The boundary will have an uncertainty of zero because the tissue with more certain boundary resolves the ambiguity.

Let us denote the expert-provided semi-automatic segmentation with $Y \in \mathbb{R}^{N \times L}$, where N is the number of voxels and L is the number of labels. Note that we can also write $Y \in \{0,1\}^{N \times L}$ because Y is a hard (0 or 1) label. For each voxel, Y is a one-hot probability vector $\mathbf{e}_k$ (which equals 1 at location k and 0 elsewhere), where k is the indicated tissue label for that voxel. We define $Y^* \in \mathbb{R}^N$ as $Y^* = \text{argmax}_{l \in [1, L]}(Y)$; in other words $Y(i) = \mathbf{e}_k \Rightarrow Y^*(i) = k$. We denote with $U \in \mathbb{R}^N$ the tissue boundary uncertainty map. U is obtained from Y by simply setting $U(i)$ to the boundary uncertainty of the tissue label for $Y(i)$. If, for example, tissue label for $Y(i)$ is caudate nuclei (CD), then $U(i) = 2$ because the boundary uncertainty for CD is two voxels. We use i^r to denote all voxels that are within a distance r from voxel i; in other words $i^r = \{k, \|k - i\| \leq r\}$. Throughout this paper, $\|.\|$ denotes the ℓ_2-norm. Also note that all voxel indices are in fact 3D indices (i.e., they have xyz elements) and, hence, $\|k - i\|$ is a distance in $\mathbb{R}^3$. However, we use single letters for indices in order to simplify the notation. Finally, we use $Y(i^r)$ to denote the "patch" of Y centered on voxel i with a radius r.

Algorithm 1: The proposed segmentation label smoothing algorithm.

Input: hard segmentation labels $Y \in \mathbb{R}^{N \times L}$,
tissue boundary uncertainty map $U \in \mathbb{R}^{N}$,
and upper bound on tissue boundary uncertainty R.
Output: smoothed segmentation labels $Y_S \in \mathbb{R}^{N \times L}$.
Initialize: $Y_S = \mathbf{0} \in \mathbb{R}^{N \times L}$.
for $i \in [1, N]$ **do**
  if $std[Y^*(i^R)] = 0$ *or* $min[U(i^R)] = 0$ **then**
    $Y_S(i) = Y(i)$;
  else
    $r_u = \min[U(i^R)]$;
    $W_k \propto \exp(-\|k - i\|/r_u) \quad \forall k \in i^{r_u}$;
    $Y_S(i)[l] = \sum_{i^{r_u}} W \odot \mathcal{P}(Y(i^{r_u}) = l) \quad \forall l \in [1, L]$;

Now, given manual segmentation labels, $Y \in \mathbb{R}^{N \times L}$, we would like to compute smoothed labels, $Y_S \in \mathbb{R}^{N \times L}$, that account for uncertain tissue boundaries. To do this for voxel i, we first consider i^R, where $R = 4$ is the upper bound of boundary uncertainty reported by our annotators for all tissues. If $Y(i^R)$ is homogeneous, that is, std[$Y^*(i^R)$] = 0, it means that voxel i is far from tissue boundaries since all voxels in i^R have the same label. Otherwise, voxel i is close to a boundary. In that case, we compute the boundary uncertainty as $r_u = \min\left[U\left(i^R\right)\right]$, i.e., the minimum of the uncertainty of the tissues in i^R. This is done following the justification provided above since the tissue with the lower uncertainty determines the uncertainty of the boundary. We then use a weighted average of tissue probabilities in i^{r_u} to compute the smoothed class probability vector for this voxel. Specifically, we use a weight matrix $W_k \propto \exp(-\|k - i\|/\tau) \quad \forall k \in i^{r_u}$, which gives higher weights to voxels that are closer to voxel i. We have found that the kernel width, τ, should depend on the patch size, which is in turn related to the boundary uncertainty. This also makes intuitive sense because for more uncertain boundaries (larger r_u) a larger kernel width should be used to achieve a higher degree of smoothing. In the experiments reported in this paper we simply set $\tau = r_u$, which we found empirically to work well. In order to ensure that the class probability vector for each voxel sums to one, W_k is normalized to sum to one. For each class label l, the smoothed probability at voxel i is then computed as:

$$Y_S(i)[l] = \sum W \odot \mathcal{P}\left(Y\left(i^{r_u}\right) = l\right), \tag{1}$$

where $\odot$ denotes element-wise (Hadamard) product and the summation is carried out over all voxels in the patch i^{r_u}. Figure 1 shows example label smoothing results generated with our proposed method. These examples show that our method smooths the boundary of each tissue/label based not only on the boundary uncertainty of that tissue but also on the boundary uncertainty of the adjoining tissues. For example, at the locations where a tissue shares a boundary with the lateral ventricles (which have a boundary uncertainty of zero voxels), the boundary becomes unambiguous and no smoothing is performed.

2.2.2. Loss function.—We use a loss function that treats certain and uncertain regions differently. We use M to denote a binary mask that shows voxels with uncertain (smoothed) labels. M is easily obtained as voxels where the maximum class probability in Y_S is not equal to one, or equivalently, as voxels whose labels are altered in the process of label smoothing. In other words:

$$(Y_s[i] \neq Y[i]) \Rightarrow M[i] = 1. \quad (2)$$

Then, our loss function is:

$$Loss\left(\hat{Y}, Y_S\right) = \sum_{M=0} \mathscr{L}\left(\hat{Y}, Y_S\right) + \sum_{M=1} \sum_{l} T^{-\mathrm{T}} \mathscr{L}\left(\hat{Y}, Y_S\right), \quad (3)$$

where $T^{-\mathrm{T}}$ is the inverse of transpose of T, $\hat{Y}$ is the segmentation map predicted by our DL model, and $\mathscr{L}$ is the base loss function, which we choose to be the cross-entropy. It has been shown that multiplication with $T^{-\mathrm{T}}$, for the uncertain boundary voxels in the second loss term above, leads to an unbiased minimizer (Patrini et al., 2017). In other words, the minimizer of the corrected loss function is the same as the minimizer obtained with clean (true) labels. Even though in this work T^{T} was non-singular, instead of $T^{-\mathrm{T}}$ we used $(T^{\mathrm{T}} + \lambda I)^{-1}$ with $\lambda = 1$ as suggested in prior works (Patrini et al., 2017) because it resulted in faster and more stable training.

We computed the label transition matrix T empirically from our training data. Specifically, after applying Algorithm 1 on the training labels, we computed T from 50 of the training images as $T_{i,j} = \Sigma_{\{Y=j \cup M=1\}} \mathscr{P}(Y_s = i)$, where summation is performed over all images. This empirical estimation is based on the standard definition of the label transition matrix. We normalized each column of T to sum to unity because it should be a left stochastic matrix. Figure 2 shows our estimated T for younger and older fetuses.

2.2.3. Implementation and experiments.—As the baseline for implementation and comparison of different methods (described below) we used the nnU-Net framework (Isensee et al., 2021). nnU-Net is considered to be the state of the art in medical image segmentation. In particular, on 53 different segmentation tasks, nnU-Net has shown that with proper selection of the training pipeline settings, standard 3D U-Net architectures (Çiçek et al., 2016) can match or outperform more elaborate network architectures. Hence, we adopt nnU-Net's network architecture (i.e., a 3D U-Net) and follow its training and inference strategies. We assess the effectiveness of our proposed methods by comparing against the methods below.

- **nnU-Net.** We followed the methods and settings in Isensee et al. (2021). In particular, we used the default loss function, which is the sum of cross-entropy and Dice.
- We tried three alternative loss functions. These included 1) **Generalized Dice** (Sudre et al., 2017) which has been proposed to improve the segmentation accuracy when the target objects are small or suffer from severe class imbalance, 2) **Focal loss** (Lin et al., 2017), which has been devised to address the extreme class imbalance by down-weighting the impact of structures that are easier to segment, and 3) improved Mean Absolute Error (**iMAE**) (Wang et al., 2019), which has been proposed for training DL models under strong label noise.

- **Training on clean labels**. In this approach, instead of training on the 272 images with noisy segmentations, we used leave-one-out cross-validation to train and test using the 22 images with highly accurate labels, which we have called "test subjects" so far. Of those 22 subjects, 8 were younger fetuses (GA < 32) and 14 were older fetuses (GA ≥ 32). Therefore, for younger fetuses each time we trained the model on 7 images and tested on the remaining image, and for older fetuses each time we trained on 13 of the images and tested on the remaining image. In this approach, each of the 22 images was used as a test image in exactly one of the experiments. Therefore, the test set for this experiment was the same as for the other methods. Since the number of training images in this approach is small, we performed this experiment both without and with transfer learning, which is a common method for dealing with limited training data (Cheplygina et al., 2019; Karimi et al., 2021). For transfer learning, we used 400 subjects from the Developing Human Connectome Project dataset (DHCP) (Hughes et al., 2017; Cordero-Grande et al., 2016) for pre-training. To the best of our knowledge, this is the most similar public dataset for the application considered in this work. It includes T2 images and tissue segmentation maps with 87 labels for newborns with GA in the range 29–45 weeks. For transfer learning, we pre-trained the network using this dataset. We then replaced the last network layer (i.e., the segmentation head) with a new head to match the number of labels considered in this work (34 for the younger fetuses and 32 for the older fetuses). We fine-tuned the pre-trained network on our data following the same leave-one-out cross-validation approach described above. We performed this separately for younger and older fetuses. Moreover, as in all experiments except for UNet++ and DeepLab mentioned below, the network architecture was the same 3D U-Net from the nnU-Net framework.
- **Standard label smoothing**. Following the standard label smoothing approach (Szegedy et al., 2016; Pereyra et al., 2017; Müller et al., 2019), we set $Y_s[i] = (1 - \alpha) Y[i] + \alpha/L$ for every voxel, except for the background voxels. We use $\alpha = 0.1$, which previous studies have shown to be a good setting (Pereyra et al., 2017; Müller et al., 2019).
- **SVLS** (Islam and Glocker, 2021). SVLS is a label smoothing method, recently proposed for medical image segmentation.
- **UNet**++. In order to also investigate the potential impact of network architecture, we compared with UNet++ (Zhou et al., 2018). This is a more elaborate nested U-Net architecture that has been proposed specifically for medical image segmentation, which claims to be better than the standard U-Net.
- **DeepLab.** DeepLab is a popular deep learning model for semantic segmentation (Chen et al., 2017). The novelty of the network architecture is the use of atrous convolutions. Furthermore, a Conditional Random Field is used to improve the resolution of the segmentation predictions.

Note that in all of the above approaches, except for UNet++ and DeepLab, we followed the same nnU-Net framework for the choice of network architecture and training settings. We

refer to Isensee et al. (2021) for the details of this framework. For UNet++, we followed the settings of the original paper (Zhou et al., 2018). Furthermore, as mentioned above, for all methods except for "Training on clean labels", we had 272 training images. For each method, we first selected a good initial learning rate using a subset of 100 images. We then trained the model using the selected initial learning rate on all 272 images. We used the "poly" learning rate decay as in Isensee et al. (2021). All training and test runs were performed using TensorFlow 1.14 under Python 3.7 on a Linux computer with an NVIDIA GeForce GTX 1080 GPU. The source code, trained model, and sample image data and segmentation labels for this work have been made publicly available at https://github.com/bchimagine/fetal_tissue_segmentation.

3. Results and Discussion

Table 1 shows the summary of the segmentation accuracy results in terms of Dice Similarity Coefficient (DSC), 95 percentile of the Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD). Our method has achieved the best results in terms of all three criteria. To determine the statistical significance of the differences, we performed paired t-tests to compare our method with every competing method in terms of these three criteria. These tests showed that, with a p-value threshold of 0.001, our method achieved significantly higher DSC and significantly lower HD95 and ASSD than all other methods, both for younger and older age groups.

In terms of all metrics, nnU-Net was the second best method after our proposed method. As we mentioned above, we used the default loss function, which is the sum of Dice and cross-entropy. Compared with this default loss function, Generalized Dice and Focal Loss performed poorly because they systematically missed one or two of the structures. That is, with Generalized Dice and Focal Loss, the network output for one or two of the labels was empty. The missed structure(s) changed with network weight initialization, but they were usually the smaller structures such as amygdala, caudate, or subthalamic nuclei. Overall, in this application with more than 30 labels we have found that loss functions based on Dice do not perform well. This may be due to the fact that the overall loss is the sum of the loss on individual labels and as the number of labels increases the relative contribution of each of the labels to the total loss becomes smaller. Since the Dice is limited to the range [0,1], the worst-case effect of a label on the total loss is $(1/L)$, where L is the number of labels. In our application with $L \approx 33$, the effect of completely missing one of the labels is only 3%. As a result, training is prone to ignoring one the labels entirely and proceed to reduce the overall loss by improving the segmentation of the other labels.

Compared with the Generalized Dice and the Focal Loss, the iMAE loss performed comparatively better and always segmented all the labels. However, it did not perform as well as nnU-Net's default loss. The iMAE loss has been proposed to strike a balance between the Mean Absolute Error (MAE) and cross-entropy. Although some studies in image classification have shown improved accuracy with MAE and iMAE when label noise is high, our results show that in the application considered in this study they do not lead to the highest segmentation accuracy. Another limitation of the iMAE loss is much longer training time, as we discuss further below.

Training on clean labels was not effective, evidently because of the much reduced number of training images. Although this approach used more accurate manual labels for training, it was limited to far fewer training images (7 or 13 training in leave-one-out cross-validation experiments, compared with 272 training images for the other methods in Table 1). Although in some applications 7–13 training images may be adequate to achieve high segmentation accuracy (Karimi et al., 2021), the application considered in this work is especially challenging due to the rapid fetal brain development and significant changes in the brain size and shape. Because of the rapid developments in the shape and complexity of structures such as the cortical gray matter, much larger numbers of training images are needed to allow the network to effectively learn these structures across the gestational age. As mentioned in the Methods section above, this experiment was performed both without and with transfer learning. As shown in Table 1, there was a consistent but small improvement in segmentation accuracy due to transfer learning. We used paired t-tests to assess the statistical significance of these differences. The tests showed a significant reduction in HD95 ($p < 0.001$) for both younger and older fetuses, although no significant differences ($p \approx 0.16 - 0.35$) in DSC or ASSD were found for either younger or older fetuses. The results presented for transfer learning in Table 1 were obtained by fine-tuning all layers of the pre-trained network. We experimented with other transfer learning approaches such as shallow fine-tuning (Tajbakhsh et al., 2016; Karimi et al., 2021) but did not achieve better results. The results of these experiments suggest that, in the application considered in this work, training with a small number of manually labeled images cannot achieve the same level of segmentation accuracy as training with a much larger number of training images with less accurate labels.

Standard label smoothing and SVLS did not work well and achieved worse accuracy metrics than nnU-Net. Standard label smoothing treats all voxels, including voxels that are far from any tissue boundary, in the same way. The SVLS method only smooths the boundary voxels, but it uses a label smoothing approach that does not take into account the actual tissue-dependent boundary uncertainties. The results obtained with these two methods shows that un-informed or spatially uniform label smoothing is not useful for segmentation with noisy labels. Finally, UNet++ and DeepLab did not perform as well as nnU-Net, confirming the arguments presented by Isensee et al. (2021) that more elaborate network architectures and post-processing operations do not necessarily lead to better results.

Figure 3 shows more detailed label-specific DSC comparison of our method with nnU-Net, which was better than the other competing methods in terms of the overall segmentation accuracy, as shown in Table 1. For structures that have separate left and right labels, we have combined the two labels into one label in order to produce a less cluttered plot. On younger fetuses, our method achieved significantly higher DSC than nnU-Net on all structures ($p < 0.001$), except for CC, where the difference was not significant. For older fetuses, our method achieved significantly higher DSC on 14 of the structures ($p < 0.001$), while on the remaining four structures (LN, TH, WM, and CSF) although the mean DSC for our method was higher, the differences were not significant. Figure 4 shows similar plots for HD95 and ASSD. For HD95, paired t-tests showed that our proposed method achieved significantly ($p < 0.001$) lower errors than nnU-Net on all structures except for LN and CR on older fetuses.

For ASSD, our method achieved significantly ($p < 0.001$) lower errors on all structures except for CR on older fetuses.

Figure 5 shows two example segmentation maps, for one younger fetus and one older fetus, predicted by our method and nnU-Net. Both our proposed method and nnU-Net succeed in segmenting different structures with good accuracy. Nonetheless, the segmentations produced by the proposed method were consistently superior. The segmentation results produced by nnU-Net often included clearly visible errors that were not present in the segmentations produced with the proposed method.

Figure 6 shows example segmentations for specific structures, which allow us to better visualize the segmentations errors and highlight under-segmentations and over-segmentations separately. Both our proposed method and nnU-Net segment most structures with good accuracy. Our method produces visibly superior segmentations on almost all structures and has less under-segmentation and less over-segmentation. Some of these structures, such as the cortical plate, have complex 3D geometries that are very difficult to segment manually. Although the additional errors in nnU-Net's output compared with our propsoed method may seem small, errors of this magnitude can make analyses which rely on consistent topology impossible and may necessitate long hours of manual correction. Our method produces better segmentations that can reduce the required manual corrections and enhance the accuracy and reproducibility of automatic analysis pipelines.

All methods considered for comparison above are based on fully convolutional networks. Further comparison with a non-DL method can be instructive and useful. To this end, we compared our proposed method with atlas-based segmentation, which is a popular classical method (Cabezas et al., 2011; Aljabar et al., 2009). We used a multi-atlas segmentation (MAS) method similar to that described in Section 2.1, without the manual refinement, to segment the 22 test images. MAS approach segmented a target image via diffeomorphic deformable registration of at least 3 atlases to the target image, followed by label fusion using probabilistic STAPLE (Akhondi-Asl and Warfield, 2013). The mean DSC of MAS segmentations on younger and older fetuses were, respectively, 0.875 ± 0.102 and 0.874 ± 0.114. These values were significantly ($p < 0.001$) lower than those achieved by the proposed method presented in Table 1. The only structure that MAS segmented slightly more accurately than the proposed method was the cerebellum on the younger fetuses (mean DSC of 0.924 for MAS versus 0.921 for the proposed method that was statistically not significant, $p = 0.30$). For all other structures, the proposed method achieved more accurate segmentation that were mostly statistically significant.

The difference between the proposed method and MAS were largest for more convoluted structures such as cortical plate (CP), subplate zone (SP), intermediate zone (IZ), ventricular zone (VZ), and white matter (WM). We show examples of the segmentations produced with the proposed method and MAS in Figure 7. For complex structures such as CP the segmentations produced by MAS show large errors. There are at least two causes for these errors. First, the registration between the atlas and the target images is never perfect, and the registration error is especially larger at the locations of thin and complex structures such as CP. Registration is difficult in fetal MRI due to stronger partial volume effects

and reduced spatial resolution. Second, there is significant inter-subject variability in the local shape of these structures, and atlases cannot fully represent this variability. Therefore, regardless of the registration error, the segmentation accuracy of atlas-based methods is limited by the fact that an atlas can only represent the “average shape” and fails to capture the inter-subject variability. Deep learning methods avoid these pitfalls because they learn the complex relationship between the image intensity maps and the target segmentation map directly from the subject training data, rather than inferring it from an atlas.

The training time for our proposed method and all other deep learning methods was 30 hours. The only exception was training with the iMAE loss that required 100 hours. The inference time for a test image was approximately 4–6 seconds, depending on the brain size with larger brains taking closer to 6 seconds. For MAS, non-rigid registration of the atlas images to the target image took approximately 12 minutes and the label fusion with the probabilistic STAPLE took approximately 13 minutes, for a total of approximately 25 minutes.

4. Conclusions

As the diagnostic quality of fetal MRI improves and its applications grow, accurate quantitative analysis methods are increasingly needed to take full advantage of this imaging technique. Training DL models to address the needs of this application is challenged by the typically low image quality and by difficulty of obtaining large labeled datasets. Our study is a step forward in improving the accuracy and reproducibility of fetal MRI analysis as it enables accurate segmentation of the brain into more than 30 relevant and important tissue compartments. To address the main challenges outlined above, we used a semi-automatic method to obtain segmentation labels on our training images. This enabled us to segment a large number of training images with reasonable annotator time. We developed a novel training method based on a label smoothing strategy that accounted for tissue boundary uncertainties. This enabled us to account for the label noise in a systematic and principled manner in our model training. Our evaluations showed that our method produced significantly more accurate segmentations than the state of the art. The improved segmentation accuracy offered by our proposed method can translate into more accurate and more reproducible quantitative assessment of fetal brain development. It can also lead to substantial reductions in the annotator time because manual segmentations are very time-consuming. Finally, our methods may be useful in many similar setting in medical image segmentation where boundary uncertainties are important.

Acknowledgments

This project was supported in part by the National Institute of Biomedical Imaging and Bioengineering, the National Institute of Neurological Disorders and Stroke, and the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (NIH) under award numbers R01EB031849, R01EB032366, R01HD109395, R01NS106030, K23NS101120, and R01NS121334; in part by the Office of the Director of the NIH under award number S10OD0250111; in part by the National Science Foundation (NSF) under award 2123061; in part by a research award from the Additional Ventures; in part by NVIDIA corporation through the Applied Research Accelerator Program and the Academic Hardware Grant Program; and in part by a Technological Innovations in Neuroscience Award from the McKnight Foundation. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF, Additional Ventures, or the McKnight Foundation.

We used anatomical MRI scans from the DHCP in a transfer learning experiment. This dataset is provided by the developing Human Connectome Project, KCL-Imperial-Oxford Consortium funded by the European Research Council under the European Union Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. [319456]. We are grateful to the families who generously supported this trial.

References

Akhondi-Asl A, Warfield SK, 2013. Simultaneous truth and performance level estimation through fusion of probabilistic segmentations. IEEE transactions on medical imaging 32, 1840–1852. [PubMed: 23744673]

Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D, 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage 46, 726–738. [PubMed: 19245840]

Bekker AJ, Goldberger J, 2016. Training deep neural-networks based on unreliable labels, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2682–2686.

Cabezas M, Oliver A, Lladó X, Freixenet J, Cuadra MB, 2011. A review of atlas-based segmentation for magnetic resonance brain images. Computer methods and programs in biomedicine 104, e158–e177. [PubMed: 21871688]

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL, 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40, 834–848. [PubMed: 28463186]

Cheplygina V, et al. , 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis 54, 280–296. [PubMed: 30959445]

Chorowski J, Jaitly N, 2016. Towards better decoding and language model integration in sequence to sequence models. arXiv preprint arXiv:1612.02695

Çiçek Ö, Abdulkadir A, et al., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Corbett-Detig J, et al. , 2011. 3d global and regional patterns of human fetal subplate growth determined in utero. Brain Structure and Function 215, 255–263. [PubMed: 21046152]

Cordero-Grande L, et al. , 2016. Sensitivity encoding for aligned multishot magnetic resonance reconstruction. IEEE Transactions on Computational Imaging 2, 266–280.

Diniz JOB, Ferreira JL, Diniz PHB, Silva AC, de Paiva AC, 2020. Esophagus segmentation from planning ct images using an atlas-based deep learning approach. Computer Methods and Programs in Biomedicine 197, 105685. [PubMed: 32798976]

Dolz J, Desrosiers C, Ayed IB, 2018. 3d fully convolutional networks for subcortical segmentation in mri: A large-scale study. NeuroImage 170, 456–470. [PubMed: 28450139]

Dou H, et al. , 2020. A deep attentive convolutional neural network for automatic cortical plate segmentation in fetal mri. IEEE transactions on medical imaging 40, 1123–1133.

Dumast P.d., et al., 2021. Segmentation of the cortical plate in fetal brain mri with a topological loss, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. Springer, pp. 200–209.

Ebner M, et al. , 2020. An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain mri. NeuroImage 206, 116324. [PubMed: 31704293]

Egaña-Ugrinovic G, et al. , 2013. Differences in cortical development assessed by fetal mri in late-onset intrauterine growth restriction. American journal of obstetrics and gynecology 209, 126–e1.

Fetit AE, et al., 2020. A deep learning approach to segmentation of the developing cortex in fetal brain mri with minimal manual labeling, in: Medical Imaging with Deep Learning, PMLR. pp. 241–261.

Fidon L, et al., 2021. Distributionally robust segmentation of abnormal fetal brain 3d mri, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. Springer, pp. 263–273.

Gholipour A, et al. , 2010. Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain mri. IEEE transactions on medical imaging 29, 1739–1758. [PubMed: 20529730]

Gholipour A, et al. , 2017. A normative spatiotemporal mri atlas of the fetal brain for automatic segmentation and analysis of early brain growth. Scientific reports 7, 1–13. [PubMed: 28127051]

Gousias IS, et al. , 2012. Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. Neuroimage 62, 1499–1509. [PubMed: 22713673]

Hosny IA, Elghawabi HS, 2010. Ultrafast mri of the fetus: an increasingly important tool in prenatal diagnosis of congenital anomalies. Magnetic resonance imaging 28, 1431–1439. [PubMed: 20850244]

Hughes EJ, et al. , 2017. A dedicated neonatal brain imaging system. Magnetic resonance in medicine 78, 794–804. [PubMed: 27643791]

Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH, 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211. [PubMed: 33288961]

Islam M, Glocker B, 2021. Spatially varying label smoothing: Capturing uncertainty from expert annotations, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 677–688.

Kainz B, et al. , 2015. Fast volume reconstruction from motion corrupted stacks of 2d slices. IEEE transactions on medical imaging 34, 1901–1913. [PubMed: 25807565]

Karimi D, Dou H, Warfield SK, Gholipour A, 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis 65, 101759. [PubMed: 32623277]

Karimi D, Samei G, Kesch C, Nir G, Salcudean SE, 2018. Prostate segmentation in mri using a convolutional neural network architecture and training strategy based on statistical shape models. International journal of computer assisted radiology and surgery 13, 1211–1219. [PubMed: 29766373]

Karimi D, Warfield SK, Gholipour A, 2021. Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. Artificial Intelligence in Medicine 116, 102078. [PubMed: 34020754]

Karimi D, Zeng Q, Mathur P, Avinash A, Mahdavi S, Spadinger I, Abolmaesumi P, Salcudean SE, 2019. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. Medical Image Analysis 57, 186 – 196. URL: http://www.sciencedirect.com/science/article/pii/S1361841519300623, doi:10.1016/j.media.2019.07.005. [PubMed: 31325722]

Khalili N, et al. , 2019. Automatic brain tissue segmentation in fetal mri using convolutional neural networks. Magnetic resonance imaging 64, 77–89. [PubMed: 31181246]

Kuklisova-Murgasova M, et al. , 2012. Reconstruction of fetal brain mri with intensity matching and complete outlier removal. Medical image analysis 16, 1550–1564. [PubMed: 22939612]

Li L, Sinclair M, Makropoulos A, Hajnal JV, David Edwards A, Kainz B, Rueckert D, Alansary A, 2021. Cas-net: Conditional atlas generation and brain segmentation for fetal mri, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. Springer, pp. 221–230.

Lin TY, Goyal P, Girshick R, He K, Dollár P, 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Liu B, et al. , 2021. The devil is in the margin: Margin-based label smoothing for network calibration. arXiv preprint arXiv:2111.15430

Makropoulos A, et al. , 2018. A review on automatic fetal and neonatal brain mri segmentation. NeuroImage 170, 231–248. [PubMed: 28666878]

Mlczoch E, et al. , 2013. Structural congenital brain disease in congenital heart disease: results from a fetal mri program. European Journal of Paediatric Neurology 17, 153–160. [PubMed: 22944287]

Müller R, et al. , 2019. When does label smoothing help? Advances in neural information processing systems 32.

Nguyen-Quynh TT, Kim SH, Do NT, 2020. Image colorization using the global scene-context style and pixel-wise semantic segmentation. IEEE Access 8, 214098–214114.

Oguz BU, Wang J, Yushkevich N, Pouch A, Gee J, Yushkevich PA, Schwartz N, Oguz I, 2018. Combining deep learning and multi-atlas label fusion for automated placenta segmentation from 3dus, in: Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis. Springer, pp. 138–148.

Patrini G, et al., 2017. Making deep neural networks robust to label noise: A loss correction approach, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1944–1952.

Payette K, Kottke R, Jakab A, 2020. Efficient multi-class fetal brain segmentation in high resolution mri reconstructions with noisy labels, in: Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Springer, pp. 295–304.

Payette K, et al. , 2021. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Scientific Data 8, 1–14. [PubMed: 33414438]

Pereyra G, et al. , 2017. Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548

Rajchl M, Lee MC, Oktay O, Kamnitsas K, Passerat-Palmbach J, Bai W, Damodaram M, Rutherford MA, Hajnal JV, Kainz B, et al. , 2016. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE transactions on medical imaging 36, 674–683. [PubMed: 27845654]

Rusiecki A, 2019. Trimmed robust loss function for training deep neural networks with label noise, in: International Conference on Artificial Intelligence and Soft Computing, Springer. pp. 215–222.

Song H, et al. , 2020. Learning from noisy labels with deep neural networks: A survey. arXiv preprint arXiv:2007.08199

Sudre CH, et al., 2017. 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. Springer, pp. 240–248.

Sukhbaatar S, et al. , 2014. Training convolutional networks with noisy labels. arXiv preprint arXiv:1406.2080

Sun L, et al. , 2019. A 3d spatially weighted network for segmentation of brain tissue from mri. IEEE transactions on medical imaging 39, 898–909. [PubMed: 31449009]

Szegedy C, et al., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

Tajbakhsh N, Jeyaseelan L, Li Q, Chiang J, Wu Z, Ding X, 2019. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. arXiv preprint arXiv:1908.10454

Tajbakhsh N, et al. , 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312. [PubMed: 26978662]

Thekumparampil KK, Khetan A, Lin Z, Oh S, 2018. Robustness of conditional gans to noisy labels. Advances in neural information processing systems 31.

Wang X, Kodirov E, Hua Y, Robertson NM, 2019. Derivative manipulation for general example weighting. arXiv preprint arXiv:1905.11233

Weisstanner C, et al. , 2015. Mri of the fetal brain. Clinical neuroradiology 25, 189–196. [PubMed: 26063004]

Yamashita Y, et al. , 1997. Mr imaging of the fetus by a haste sequence. AJR. American journal of roentgenology 168, 513–519. [PubMed: 9016238]

Zeng Q, Samei G, Karimi D, Kesch C, Mahdavi SS, Abolmaesumi P, Salcudean SE, 2018. Prostate segmentation in transrectal ultrasound using magnetic resonance imaging priors. International journal of computer assisted radiology and surgery 13, 749–757. [PubMed: 29589259]

Zhang Z, Sabuncu M, 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems 31.

Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J, 2018. Unet++: A nested u-net architecture for medical image segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, pp. 3–11.

Our aim is automatic fetal brain tissue segmentation in MRI.

We generate noisy labels on a larger number of images using an atlas-based method.

We develop methods for training a deep learning model with these noisy labels.

On manually labeled images, our method achieves superior segmentation accuracy.

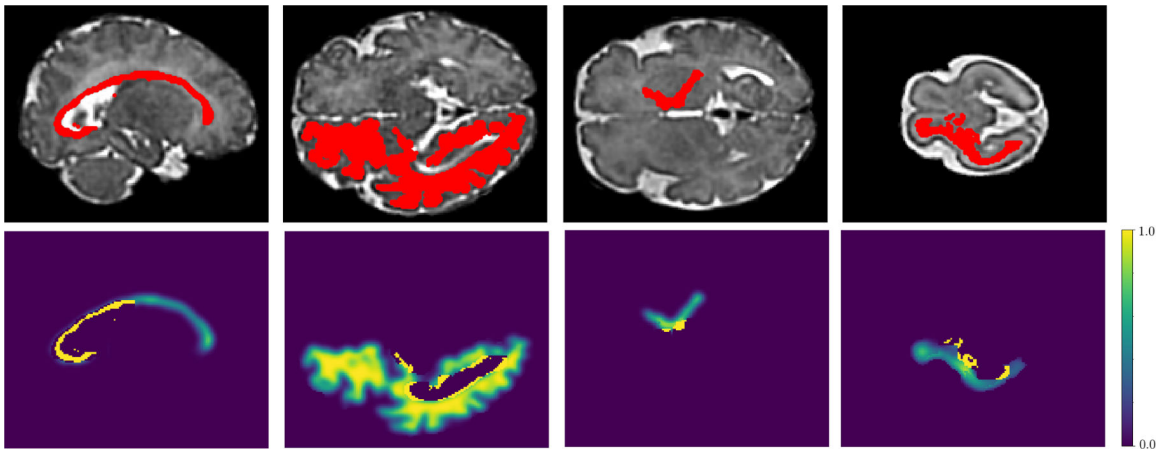

Fig. 1.

Examples of label smoothing performed by our proposed method (Algorithm 1). Note that for each image our algorithm performs the smoothing simultaneously on all labels, but here we show a single label in each example for better visualization. In each of the examples presented here, the top image shows the original label obtained via multi-atlas segmentation followed by manual correction of large errors, super-imposed on the T2 image. The lower images show the smoothed label.

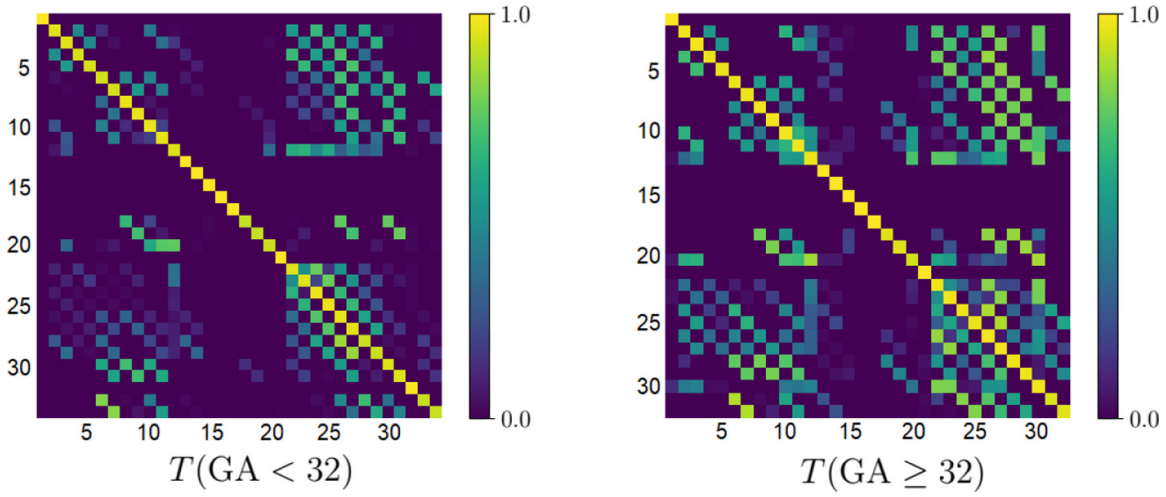

Fig. 2.

Label transition matrices, T, for the younger (left) and older (right) age groups. We have used a log transformation, $\log(T + 0.001)$, to display these matrices in order to better highlight the smaller off-diagonal elements. Label numbers have been shown on the rows and columns of the matrices. For younger fetuses the label names are as follows (acronyms have been defined in the text in Section 2.1). 1: background, 2: HP left, 3: HP right, 4: AM left, 5: AM right, 6: CD left, 7: CD right, 8: LN left, 9: LN right, 10: TH left, 11: TH right, 12: CC, 13: LV left, 14: LV right, 15: ST, 16: CR left, 17: CR right, 18: SN left, 19: SN right, 20: HC, 21: FN, 22: CP left, 23: CP right, 24: SP left, 25: SP right, 26: IZ left, 27: IZ right, 28: VZ left, 29: VZ right, 30: IC left, 31: IC right, 32: CSF, 33: GE left, 34: GE right. For older fetuses, the first 23 labels are the same as those for younger fetuses, and the remaining labels are as follows. 24: VZ left, 25: VZ right, 26: WM left, 27: WM right, 28: IC left, 29: IC right, 30: CSF, 31: GE left, 32: GE right.

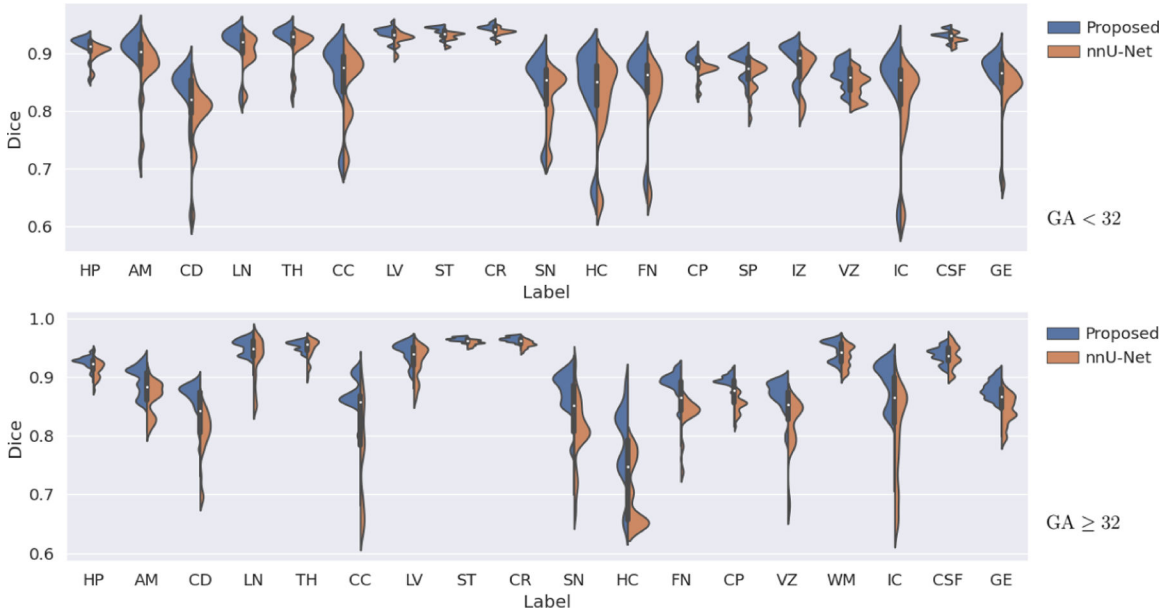

Fig. 3.

Comparison of our method and nnU-Net in terms of DSC for different structures on younger (top) and older (bottom) fetuses.

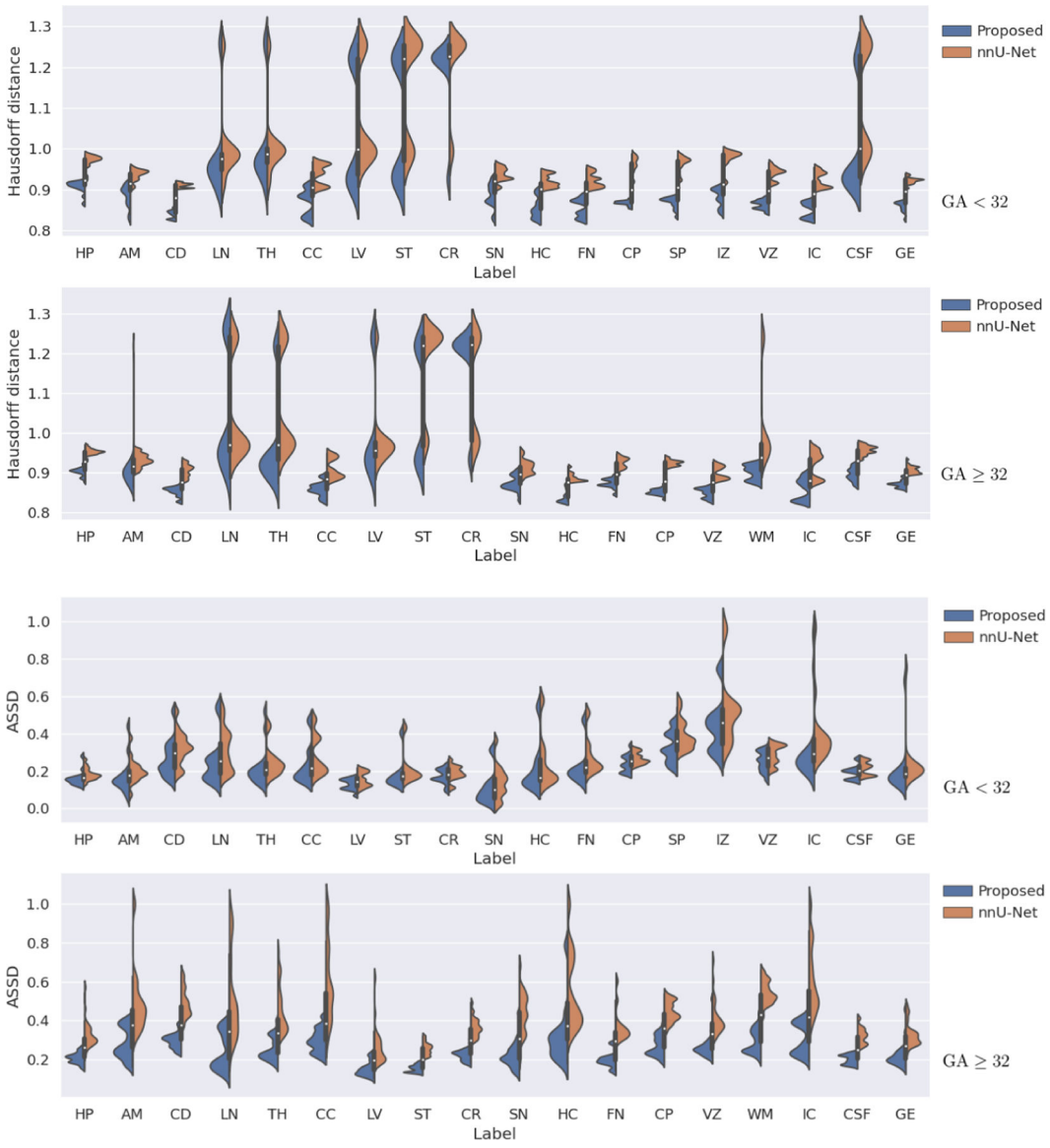

Fig. 4.

Comparison of our method and nnU-Net in terms of HD95 (the top two plots) and ASSD (the bottom two plots) on younger and older fetuses. The GA group for each plot is shown on the right of that plot.

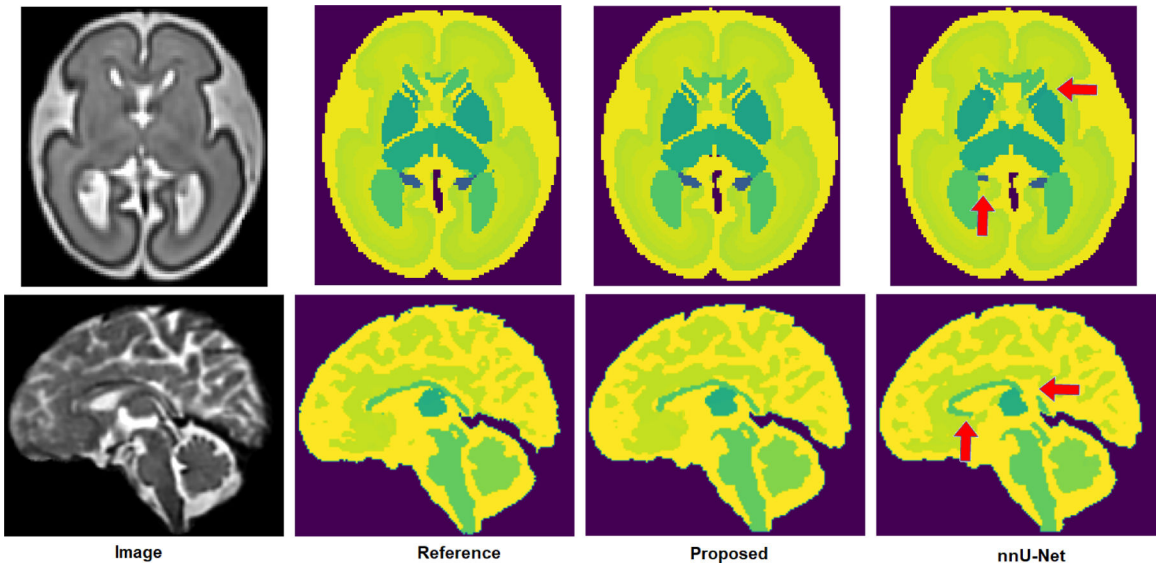

Fig. 5.

Two example segmentation maps predicted by our proposed method and nnU-Net. Red arrows point to some of nnU-Net's segmentation errors. The top example is for a younger fetus (GA= 26.7 weeks), whereas the bottom example is for an older fetus (GA= 38 weeks).

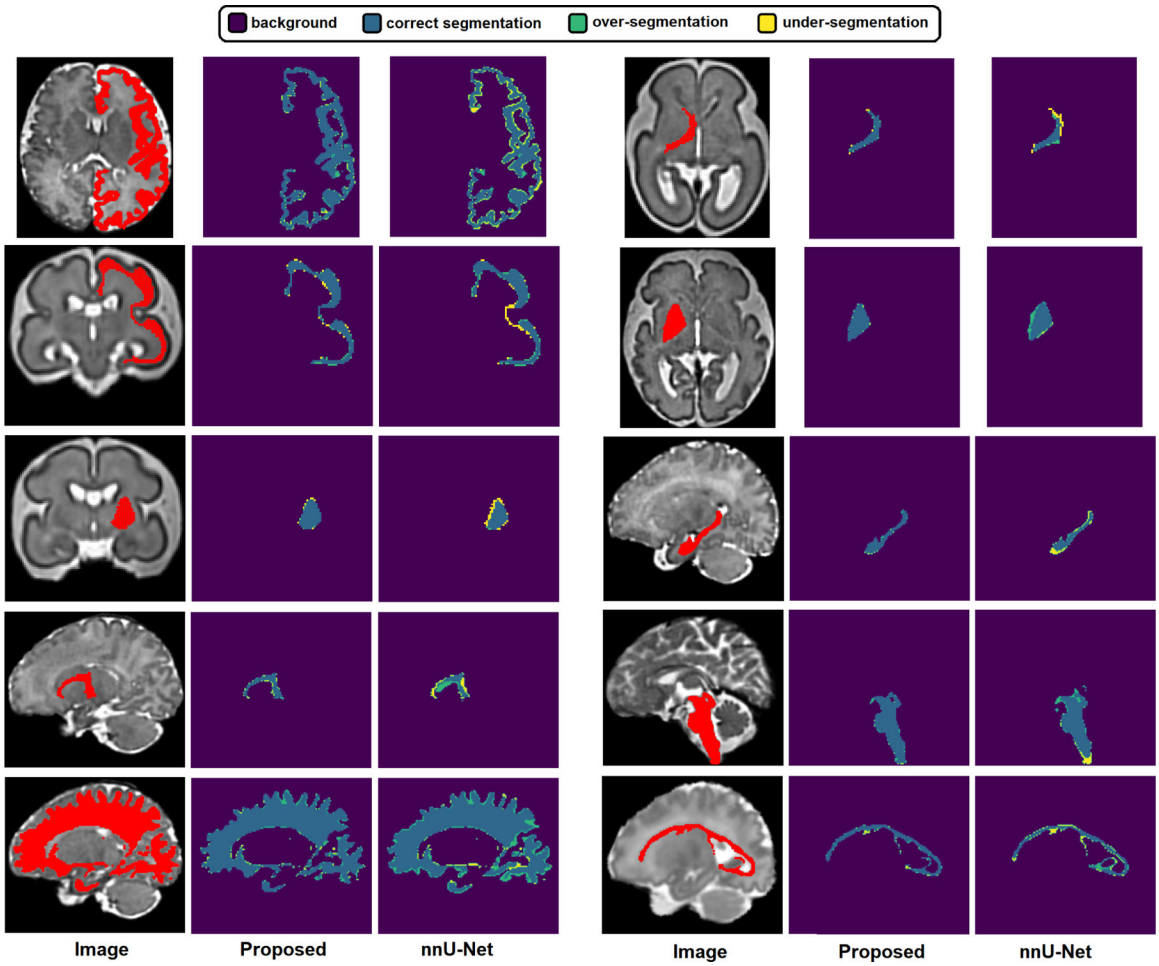

Fig. 6.

Example segmentations predicted by our proposed method and nnU-Net. Each of the ten examples shows one isolated structure and highlights correct segmentations, over-segmentations and under-segmentations.

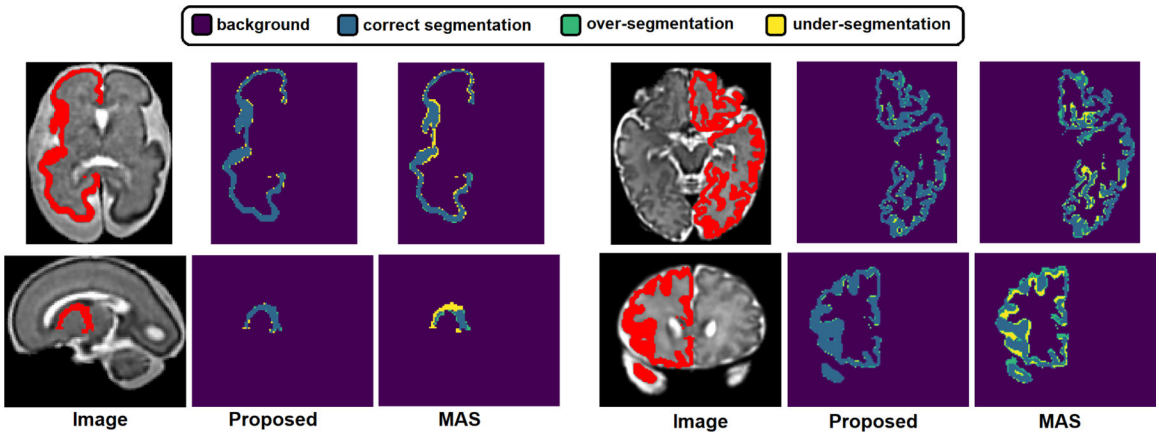

Fig. 7.

Example segmentations predicted by our proposed method and multi-atlas segmentation (MAS). Each of the four examples in this figure shows one isolated structure and highlights correct segmentations, over-segmentations and under-segmentations.

Table 1.

Segmentation accuracy metrics presented separately for younger and older fetuses. The metrics were computed separately for each label; this table presents mean ± standard deviation over all labels. Best results for each metric are in bold. We used paired t-tests to compare our proposed method with every other method. Asterisks in this table denote significantly better results for the proposed method than all other methods

(at a significant threshold of $p < 0.001$). In the Method column in this table, T.L. stands for transfer learning.

Dataset	Method	DSC	HD95 (mm)	ASSD (mm)
Younger fetuses	nnU-Net	0.872 ± 0.063	0.99 ± 0.11	0.26 ± 0.14
	Generalized Dice	0.845 ± 0.087	1.09 ± 0.12	0.32 ± 0.26
	Focal loss	0.839 ± 0.080	1.15 ± 1.20	0.30 ± 0.19
	iMAE	0.865 ± 0.075	1.06 ± 0.15	0.26 ± 0.17
	Training on clean labels (without T.L.)	0.863 ± 0.068	1.09 ± 0.14	0.28 ± 0.16
	Training on clean labels (with T.L.)	0.866 ± 0.062	1.03 ± 0.13	0.27 ± 0.17
	Standard label smoothing	0.833 ± 0.084	1.08 ± 0.17	0.34 ± 0.21
	SVLS	0.843 ± 0.074	1.07 ± 0.17	0.30 ± 0.18
	DeepLab	0.851 ± 0.072	1.11 ± 0.15	0.30 ± 0.15
	UNet++	0.866 ± 0.060	1.02 ± 0.14	0.27 ± 0.15
	Proposed method	**0.893 ± 0.066***	**0.94 ± 0.13***	**0.23 ± 0.13***
Older fetuses	nnU-Net	0.896 ± 0.066	0.98 ± 0.11	0.36 ± 0.12
	Generalized Dice	0.866 ± 0.070	1.16 ± 0.11	0.46 ± 0.15
	Focal loss	0.861 ± 0.068	1.16 ± 0.16	0.42 ± 0.16
	iMAE	0.880 ± 0.064	1.09 ± 0.17	0.41 ± 0.20
	Training on clean labels (without T.L.)	0.877 ± 0.073	1.12 ± 0.14	0.40 ± 0.18
	Training on clean labels (with T.L.)	0.880 ± 0.070	1.04 ± 0.15	0.40 ± 0.20
	Standard label smoothing	0.853 ± 0.071	1.16 ± 0.12	0.39 ± 0.23
	SVLS	0.856 ± 0.077	1.10 ± 0.13	0.37 ± 0.27
	DeepLab	0.865 ± 0.074	1.21 ± 0.19	0.43 ± 0.26
	UNet++	0.885 ± 0.070	1.08 ± 0.16	0.38 ± 0.23
	Proposed method	**0.916 ± 0.059***	**0.94 ± 0.13***	**0.25 ± 0.09***