Optimizing Segmentation of Neonatal Brain MRIs with Partially Annotated Multi-Label Data

Dariia Kucheruk[1,2,3] DARIIA.KUCHERUK@MAIL.UTORONTO.CA
Sam Osia[2,4] SAM.OSIA@SICKKIDS.CA
Pouria Mashouri[2] POURIA.MASHOURI@UHN.CA
Elizaveta Rybnikova[1,3] JAN.RYBNIKOVA@MAIL.UTORONTO.CA
Sergey Protserov[1,2,3] S.PROTSEROV@MAIL.UTORONTO.CA
Jaryd Hunter[2] JARYD.HUNTER@UHN.CA
Maksym Muzychenko[1,3] MAKSYM.MUZYCHENKO@MAIL.UTORONTO.CA
Ting Guo[4] JESSIE.GUO@SICKKIDS.CA
Michael Brudno[1,2,3,4] BRUDNO@CS.TORONTO.EDU
1. *University of Toronto*
2. *University Health Network*
3. *Vector Institute for Artificial Intelligence*
4. *The Hospital for Sick Children*

Abstract

Accurate assessment of the developing brain is important for research and clinical applications, and manual segmentation of brain MRIs is a painstaking and expensive process. We introduce the first method for neonatal brain MRI segmentation that simultaneously leverages fully and partially labeled data within a multi-label segmentation framework. Our method improves accuracy and efficiency by utilizing all available supervision—even when only coarse or incomplete annotations are present—enabling the model to learn both detailed and high-level brain structures from heterogeneous data. We validate our method on scans from the Developing Human Connectome Project (dHCP) acquired at both preterm and term gestational ages. Our approach demonstrates more accurate and robust segmentation compared to standard supervised and semi-supervised models trained with equivalent data. The results showed an improvement in predictions of predominantly unannotated labels in the training set when combined with labels of relevant "super-classes". Further experiments with semi-supervised loss functions demonstrated that limited but reliable supervision is more effective than using noisy labels. Our work presents evidence that it is possible to build robust medical image segmentation models with only a small amount of fully labeled training data. Our code is available at https://github.com/dkucheru/Brain-Segmentation-MLHC-2025.git

1. Introduction

Magnetic Resonance Imaging (MRI) is a valuable medical imaging technique that provides a 3D view of brain structures without the use of radiation. However, analysis of a brain MRI

scan often requires accurate segmentation of anatomical structures. This labor-demanding and time-consuming work typically needs to be done by a qualified radiologist or neuroscientist, and affects the timeliness of provided care and cost of creation of large-scale datasets. Building automated approaches for brain MRI segmentation would thus help reduce costs and improve care.

Multiple approaches of automated brain segmentation have been proposed (Prastawa et al., 2005; Jenkinson et al., 2012; Ashburner, 2012; Fischl, 2012), nevertheless, despite their success in classifying adult brain tissue, their effectiveness drops in segmenting infant brains, which undergo rapid development and have less contrast among neural tissues. Deep learning techniques for segmenting children's brains are available (Fetit et al., 2020; Zöllei et al., 2020; Liu et al., 2021), but are more suitable for specific age groups.

The accuracy of supervised segmentation methods is dependent on the availability and quality of labeled training datasets. However, neonatal brain MRIs are typically segmented for a limited number of cases in research studies, limiting their overall number. Moreover, the available neonatal brain scans are often segmented at varying levels of detail. For example, the brainstem consists of several sub-regions, including midbrain, pons, and medulla, and some obtained scans may have only the annotation of the whole brainstem, while others may label one (or more) of the sub-regions.

Taken together, neonatal MRI data is often limited in quantity and segmentation detail and is highly variable between hospitals due to differences in scanner hardware and acquisition protocols. This creates a need for training methods that can learn effectively from limited, partially labeled, and diverse datasets. Addressing these challenges is essential for developing segmentation models that generalize well across diverse clinical environments.

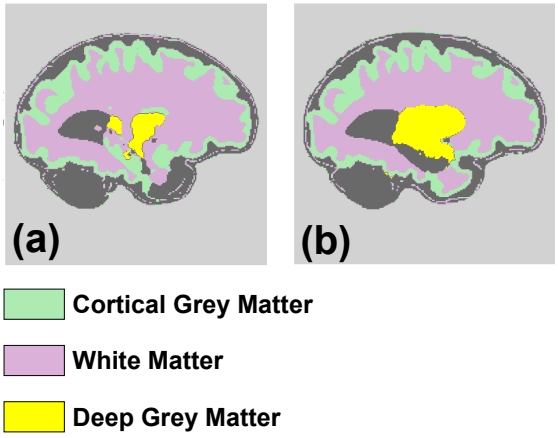


Figure 1: Comparison of segmentation results from two models trained on the same number of labels for Cortical Grey Matter, White Matter, and Deep Grey Matter. **(a)** Prediction from the fully labeled baseline. **(b)** Prediction from **our method**, which shows visibly **improved accuracy**.

In this work we propose a method that enhances the accuracy and efficiency of neonatal brain MRI segmentation by leveraging all available information from partially annotated data. Recently the use of partially labeled data has been explored in the context of classification and multi-class segmentation tasks. However, the use of **partially labeled data** in the context of **multi-label segmentation** tasks remains **largely unexplored.** Inspired by works in classification domain (Cole et al., 2021; Jo et al., 2023) we develop an approach to address the common challenge in neonatal MRI scans, where only coarse segmentations are often available. Our method allows the

training process to effectively incorporate all available scans, including those with certain features missing or only large features identified.

To evaluate the effectiveness of our pipeline, we used two subsets from the open-source Developing Human Connectome Project (dHCP) dataset (Hughes et al., 2017). Specifically, we selected MRI scans acquired at two distinct developmental stages: preterm (between 28–36 weeks postmenstrual age) and at-term (approximately 37–44 weeks postmenstrual age). From these, we sampled training datasets with different fractions of partially labeled data, and trained our model, as well as classical supervised and semi-supervised models which used equal amounts of fully labeled data as our model. The experiments show that our approach outperforms both of these approaches in terms of segmentation accuracy, while utilizing fewer computational resources than the semi-supervised approach. Biggest improvements were achieved on scans of very preterm neonates, whose brains are typically harder to annotate.

Additionally, we developed and evaluated several loss functions, comparing their performance with our proposed method. Interestingly, our results show that it is better to avoid introducing any potential labeling errors through pseudo-labeling, rather than relying on pseudo-labeled data that may degrade model performance.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work illustrates that it is possible to build robust medical image segmentation models with only a small portion of fully labeled training data. We introduce a method for effectively leveraging partially labeled data in multi-label segmentation tasks, where annotations vary in granularity. A key insight is that models trained solely on available true labels using Ignore Unobserved Child (IUC) loss function—despite inconsistencies in labeling detail—outperform those that incorporate pseudo-labels generated during training, suggesting that introducing noisy labels may harm performance more than using limited but reliable supervision. Finally, our approach is architecture-agnostic and can be extended beyond U-Net models and to other domains where multi-label partial segmentation annotations is a challenge.

We hope the community will find our method and insights useful for advancing medical image segmentation in settings where high-quality, fully labeled data is limited, particularly in neonatal brain imaging and similar healthcare applications.

2. Related Work

Semantic image segmentation is a widely studied area in computer vision, with important applications in healthcare. In neonatal care, automated brain MRI segmentation can support early diagnosis, which is especially valuable since detecting conditions early in life can greatly improve outcomes. However, working with neonatal data is challenging because the amount of data is often small, and the available labels vary in detail and completeness. These limitations make it difficult to train accurate models. We focus on addressing these challenges and improving segmentation performance under limited and partially labeled data.

2.1. Learning from Partially Labeled Data

A number of influential studies have explored the challenge of training models on scattered or incomplete data, introducing a range of innovative approaches—particularly in the context of computer vision tasks.

Among these, several works in image classification have proposed strategies to effectively learn from partially labeled or noisy datasets. For example, Zhou et al. (2022) demonstrated strong performance in single-positive multi-label classification by leveraging partial annotations. Other approaches have extended multi-label classification to settings with unknown hierarchies and incomplete labels (Jo et al., 2023; Wang et al., 2023). Ben-Baruch et al. (2022) introduced a class-aware selective loss, which models the probabilistic distribution of labels while accounting for label likelihoods specific to each input. While these methods have proven effective for classification, they are not directly applicable to image segmentation, which presents additional spatial and structural challenges.

Some segmentation methods have explored the use of training data with varying levels of annotation. For instance, Reiss et al. (2021) proposed a framework that leverages noisy annotations by applying patch-wise supervision where multiple category associations are allowed at intermediate network layers. Similarly, Valabrègue et al. (2023) demonstrated that segmentation models trained solely on synthetic MRI data—generated from a limited set of labeled neonatal brain scans—can generalize effectively across imaging contrasts. While these approaches show that high performance can be achieved without large fully labeled datasets, they focus on multi-class segmentation. The multi-label segmentation scenarios, where each pixel may belong to multiple classes simultaneously, remain unaddressed.

2.2. The Gap

Semantic segmentation has been carried out within the context of multi-label learning, employing convolutional neural networks in map segmentation (Davies, 2022) and object detection in natural environments (Jordan, 2018). One study (Lempart et al., 2022) integrated the U-NET architecture and multi-label approach. Multi-label image segmentation was also enhanced with a framework based on hierarchical structures of segmented classes (Li et al., 2022). All these studies assume that each training example is fully labeled, while real-world MRI datasets may be missing annotations of specific brain regions.

Recent advances in segmentation, such as the Segment Anything Model (SAM) (Kirillov et al., 2023), have demonstrated strong performance across diverse imaging domains, including medical imaging (Ma et al., 2024). However, these models rely on user-provided or simulated prompts (e.g., points, boxes, or masks) to guide segmentation, making them less suitable for settings where autonomous operation is essential. This dependence on prompting limits their applicability in scenarios like neonatal MRI, where manual input is impractical or unavailable at inference time.

Fully autonomous multi-label semantic segmentation with partially annotated training data remains an underexplored area of research.

Building on the ideas of Multi-Label Learning from Single Positive Labels (Cole et al., 2021), we propose a method that enhances both the precision and computational efficiency of brain MRI segmentation. Our approach fully utilizes the information present in partially

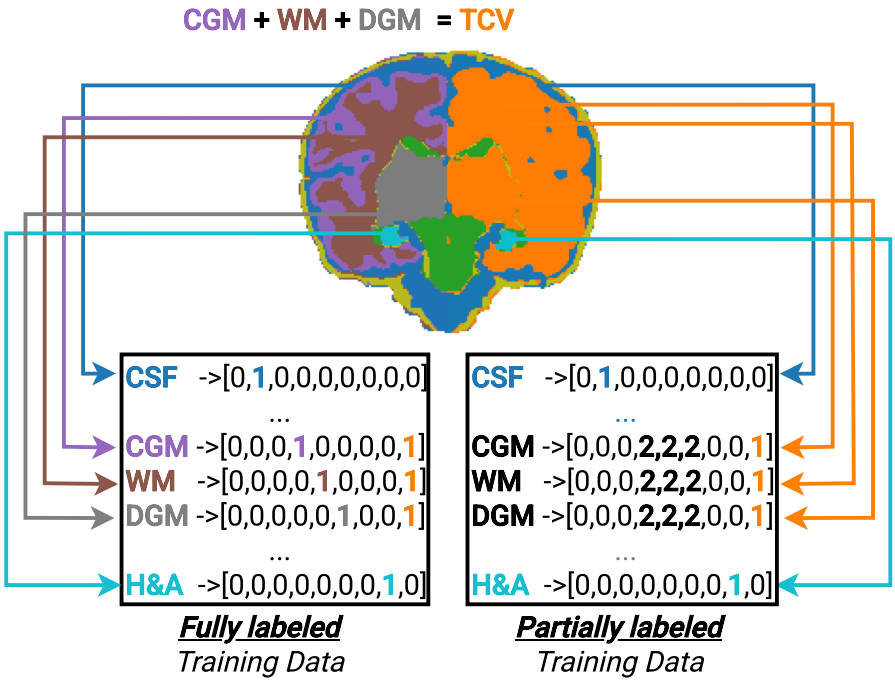


Figure 2: Illustration of multi-label assignment for fully and partially labeled training data. In this illustrative example, there are 9 anatomical regions in total. To preserve clarity and visual appeal of the diagram, only a subset is shown. The main focus is on CGM (Cortical Gray Matter), WM (White Matter), and DGM (Deep Gray Matter) - sub-regions of the superclass Total Cerebrum Volume (TCV). In the partially labeled case (right), only the segmentation for superclass (TCV) is available. The indices for the three sub-regions are marked as unobserved (set to 2), while the TCV super-label remains observed (set to 1). Regions not part of TCV—including CSF (Cerebrospinal Fluid) and H&A (Hippocampus & Amygdala), among others—remain fully labeled in both cases to illustrate unaffected labels.

labeled datasets, enabling effective use of scans where some structures are absent or only prominent regions are annotated.

3. Methods

The key contribution of our work lies within the ability of our model to incorporate partially and fully annotated multi-label segmentations into the training process. Full MRI segmentations carry detailed labels of brain regions, however partial segmentation may carry more 'high-level' super-labels. For example, fully segmented MRI scan will have Cortical Gray Matter(CGM), Deep Gray Matter(DM) and White Matter(WM) labels, whilst partially segmented MRI scan will have a broader super-label Total Cerebrum Volume(TCV) instead of

the three aforementioned sub-labels. To learn from all the available data, we create super-labels for fully segmented scans, whilst modelling unannotated sub-regions from partially labeled data as *missing labels*.

To achieve this, all of the available data has to be changed from multi-class to multi-label, so that some pixels could be assigned to Total Cerebrum Volume (TCV) and one of its sub-regions simultaneously by the segmentation method. For fully labeled datasets this is achieved by a step of one hot encoding followed by a simple step of broader category label assignment shown on the left side of the diagram (see Fig. 2).

In the classic multi-label segmentation setting, each pixel is assigned a binary vector x, where $x[i] = 1$ means that class at index i is relevant to the current pixel, and $x[i] = 0$ for irrelevant classes. As seen in previous works (Cole et al., 2021) we process partially labeled segmentation masks by marking missing sub-regions as 'unknown'. Following this idea, if it is unknown whether the pixel belongs to a certain class, the $x[i]$ will get assigned an arbitrary value used for the unknown sub-label ($x[i] = 2$ in Fig. 2). The right side of the diagram on Fig. 2 shows how the method uses the available super-label TCV for training when the sub-regions are not annotated.

A straightforward strategy for handling missing labels is to exclude their contributions from the binary cross-entropy (BCE) loss—an approach referred to as Ignore Unobserved (IU) loss by Cole et al. (2021). We build on this idea by introducing a novel application of IU in hierarchical settings, where parent labels are known but child labels may be missing. Our proposed Ignore Unobserved Child (IUC) loss leverages the consistent presence of parent classes across all scans to learn their boundaries, while also enabling the model to refine the segmentation of child regions when they are available during training. Crucially, IUC introduces no label noise, ensuring stable learning even with partially annotated data:

$$\hat{\mathcal{L}}_{\text{IUC}}(\mathbf{x}, \mathbf{y}) = -\frac{1}{|N|}\left[\sum_{l \in L}\left(x_l \log(y_l) + (1 - x_l)\log(1 - y_l)\right) + \gamma \sum_{c \in C}\left(x_c \log(y_c) + (1 - x_c)\log(1 - y_c)\right)\right],$$

where $\boldsymbol{L}$ is a set of observed larger classes, $\boldsymbol{C}$ is a set of children classes, $\boldsymbol{N}$ is the number of classes, $\boldsymbol{x}$ is a vector of original labels for the current pixel ($\mathrm{x} \in \{0, 1, 2\}^L$), $\boldsymbol{y}$ is a vector of class probabilities predicted by the model for some input pixel, and γ:

$$\gamma = \begin{cases} 1 & \text{if } x_c \in \{0, 1\} \\ 0 & \text{if } x_c \in \{2\} \end{cases}$$

3.1. Mathematical intuition of hierarchical supervision

In pixel-wise segmentation models, a shared feature vector $z_{ij} \in \mathbb{R}^d$ is computed for each pixel (i, j) to make class-specific predictions by applying independent linear classifiers $w_k \in \mathbb{R}^d$ to the shared feature vector, followed by a sigmoid activation: $\hat{y_{ijk}} = \sigma(w_k^\top z_{ij})$, where

$\hat{y_{ijk}} \in (0,1)$ – predicted probability that pixel (i,j) belongs to class k, and $\sigma(\cdot)$ – sigmoid function.

During our training, supervision may be partial—e.g., only a superclass $k_1$is labeled. The loss updates both the classifier $w_{k_1}$and the shared representation z_{ij}, which also affects predictions for other classes, such as a related subclass k_2.

As a result, the model can implicitly capture hierarchical relationships among classes through shared representations and selective supervision. This emerges naturally from the shared parameterization without explicit hierarchy constraints in the loss or architecture.

To sum up, IUC stands out by learning coarse structures from partial labels, refining fine details from limited annotations, and preserving anatomical hierarchy without relying on noisy heuristics.

3.2. Preferred Method & Loss Variants

Our key contribution is a training pipeline that uses Ignore Unobserved Child (IUC) loss and relies solely on available true labels. Unlike approaches that incorporate pseudo-labels generated during training, this design avoids the risk of introducing noisy labels, which can degrade performance. To support this design choice, we introduce three additional loss variants – HUC, WEIGHT, and WNORM – that incorporate heuristics or pseudo-labels. These variants are intended to examine how such additional supervision signals may affect performance by introducing potentially harmful noise.

As a more advanced loss formulation, we draw inspiration from the Hierarchy-aware Unobserved Positive BCE loss Modification proposed by Jo et al. (2023), originally designed for classification tasks where each training sample is annotated with only a single positive class, and all other labels are treated as missing. The primary contribution of that work lies in the automatic extraction of hierarchical structure from data; however, we were particularly motivated by the way hierarchical relationships were leveraged within the loss function.

In our work, we propose a significantly modified version of this idea, tailored for the semantic segmentation setting, which we term Hierarchy-aware Unobserved Child loss (HUC).

$$\hat{\mathcal{L}}_{\text{HUC}}(\mathbf{x},\mathbf{y}) = -\frac{1}{|N|}\left[\sum_{o\in O}\left(x_o \log(y_o) + (1-x_o)\log(1-y_o)\right) + \sum_{u\in U}\left(\delta_{u|o}\log(y_u) + (1-\delta_{u|o})\log(1-y_u)\right)\right]$$

where $\boldsymbol{O}$ is a set of observed classes, $\boldsymbol{U}$ is a set of unobserved classes, and $\mathbf{O} \cup \mathbf{U} = \mathbf{N}$

Unlike the original setup, our training data contains multiple positive and negative classes per pixel. Therefore, we directly compute the binary cross-entropy (BCE) loss for all available positive and negative labels, including their corresponding parent classes in the hierarchy.

For the remaining, unobserved child classes, we utilize an indicator-based pseudo-labeling strategy $\delta_{u|o}$:

$$\delta_{u|o} = \begin{cases} 1 & \text{if } u \in \text{Children}(o),\ x_o = 1,\ \text{and } y_u = \max\limits_{u' \in U} y_{u'} \\ 0 & \text{otherwise} \end{cases}.$$

Specifically, for each group of sibling classes under a common parent, we assume a positive pseudo-label for class u if it has the highest model prediction score within its group. All other sibling classes are treated as negative, and we compute BCE using these pseudo-labels.

We propose $\hat{\mathcal{L}}_{\text{WEIGHT}}$, a variant of $\hat{\mathcal{L}}_{\text{HUC}}$ that replaces hard pseudo-labels (0 or 1) for unobserved classes with the model's own predictions (y_{u}) in the BCE loss. This soft supervision provides a smoother training signal and encourages the model to self-regularize.

$$\hat{\mathcal{L}}_{\text{WEIGHT}}(\mathbf{x}, \mathbf{y}) = -\frac{1}{|N|}\left[\sum_{o \in O}\left(x_o \log(y_o) + (1 - x_o)\log(1 - y_o)\right) + \sum_{u \in U}\left(y_u \log(y_u) + (1 - y_u)\log(1 - y_u)\right)\right]$$

As a further refinement, we normalize the predictions for unobserved classes using L1 normalization over the corresponding child classes, ensuring their values sum to 1.

$$\hat{\mathcal{L}}_{\text{WNORM}}(\mathbf{x}, \mathbf{y}) = -\frac{1}{|N|}\left[\sum_{o \in O}\left(x_o \log(y_o) + (1 - x_o)\log(1 - y_o)\right) + \sum_{u \in U}\left(y'_u \log(y_u) + (1 - y'_u)\log(1 - y_u)\right)\right],$$

where given $u \in \text{Children}(o),\ x_o = 1$ it is true that $\sum_{u \in \text{Children}(o)} y'_u = 1$.

This way our proposed loss functions allow the training process to benefit from the high-level information present in the partially labeled data and simultaneously learn more detailed segmentations from fully labeled data.

4. Cohort

The Developing Human Connectome Project (dHCP) Dataset is the third release of open-access neonatal brain MRI data (Hughes et al., 2017). In this paper, we included 190 neonates born at 32 weeks of gestation or less from the dHCP dataset, of which 96 had both preterm MRI scans and at-term age MRI scans, 94 had only preterm MRI scans, and other 94 had only at-term age MRI scans.

The models in this work were trained with both preterm and at-term T2-weighted MRI scans from the dHCP dataset. The validation and test sets were exclusively preterm and term scans as well.

5. Results on Real Data

5.1. Evaluation Approach

We chose a U-NET architecture Ronneberger et al. (2015) to segment brain MRIs of neonates. Our U-Net has 5 encoder and 5 decoder blocks, each encoder block has 2 convolution layers and max pooling; each decoder block has 2 convolution layers, transposed convolution for upsampling, and a skip connection. This is a standard architecture that presents as sufficiently deep to handle the complexity of the segmentation task.

Each model was trained for 200 epochs with an early stopping of 30 epochs and Adam optimizer with a learning rate of 5×10^{-3}.

Each original 3D scan (256,256,256) was broken down into 2D slices of shape (256,256), and the model was trained on each slice separately. The data was split into training set (80%), validation set(10%), and test set(10%).

To evaluate our method in a semi-supervised setting, we modified the training set to contain both fully labeled (FL) and partially labeled (PL) data. Specifically, we created partially labeled examples by masking out three sub-labels—deep grey matter (DGM), white matter (WM), and cortical grey matter (CGM)—within the total cerebral volume (TCV) super-label. These sub-labels were marked as 'unknown' for a subset of the training data, simulating partial supervision (see Fig.2).

To study how performance varies with the amount of fully labeled data, we generated three experimental configurations:

- 2% of the training set fully labeled and 98% partially labeled,
- 5% fully labeled and 95% partially labeled,
- 10% fully labeled and 90% partially labeled.

To ensure the robustness and reliability of our results, we performed **five independent rounds of random selection** of fully labeled patient IDs for each configuration. That is, for each target proportion (2%, 5%, and 10%), we randomly sampled 2%, 5%, or 10% of patient IDs from the training set to remain fully labeled, while the remaining patient IDs were converted to partially labeled format (as illustrated on the right side of Fig. 2). This gave us:

- 5 distinct versions of the dataset with 2% FL and 98% PL,
- 5 versions with 5% FL and 95% PL,
- 5 versions with 10% FL and 90% PL.

This approach allows us to report averaged results across multiple random splits, making our evaluation more reliable and representative of general performance across different possible subsets of fully labeled data.

We trained three types of models in our experiments:

5.1.1. Fully Supervised Baseline

First, we trained a classic supervised model that only used fully labeled data during training. We ran 3 experiments on 2%, 5%, and 10% of the training data that remained fully labeled. To illustrate the gap between all of the models and best possible theoretical performance we also trained a fully labeled model with 100% of the data.

Overall 16 models were trained in fully supervised setting : 3 fractions (2%, 5%, 10%) x 5 rounds of random subsets, plus the 16th model that utilized all training data (100%) as fully labeled.

5.1.2. Semi-Supervised Baseline

As noted in the related works section, there is currently no open-source baseline tailored for multi-label semantic segmentation with partially annotated training data. To address this gap, we designed a semi-supervised baseline inspired by common practices in the field.

An ideal baseline would handle multi-label segmentation while also making use of partial annotations. Since no such method was available, we turned to established techniques in semi-supervised learning, specifically pseudo-labeling, and extended them to our setting using anatomical priors (Zheng et al., 2019).

Our baseline follows a simple two-stage approach. First, we trained a supervised model on a small portion of the training set that was fully labeled (2%, 5%, or 10%). This model was then used to generate pseudo-labels for the remaining partially labeled data (98%, 95%, or 90%). To incorporate available supervision from the partial annotations, we applied a correction step: each predicted sub-label was constrained not to extend beyond its corresponding known super-label. An example of this correction process is shown in Fig. 3.

This baseline setup provides a meaningful point of comparison for our method by combining established ideas from semi-supervised learning and domain-specific prior knowledge.

In total, 15 models were trained in the semi-supervised setting: 3 fractions (2%/98%, 5%/95%, 10%/90%) x 5 rounds of random subsets. The semi-supervised method converged after one iteration.

5.1.3. Our Method

To evaluate our method, which aims to leverage high-level information from partially labeled data while simultaneously learning detailed segmentations from fully labeled data, we compared several loss functions proposed in this paper:$\hat{\mathcal{L}}_{\text{IUC}}$; $\hat{\mathcal{L}}_{\text{HUC}}$; $\hat{\mathcal{L}}_{\text{WEIGHT}}$; $\hat{\mathcal{L}}_{\text{WNORM}}$.

For the $\hat{\mathcal{L}}_{\text{IUC}}$, we trained 15 models, corresponding to three fully/partially labeled fractions (2%/98%, 5%/95%, and 10%/90%) and five random subset selections per fraction. For the other three loss functions, we trained 10 models each, using the 2%/98% and 5%/95% fractions with five different random subsets.

To evaluate the performance of the models, we analyzed the Dice score and Hausdorff Distance of the segmentations predicted by the models compared to the ground truth values. We average predictions across five randomly selected fractions (2%, 5%, and 10%) of subjects.

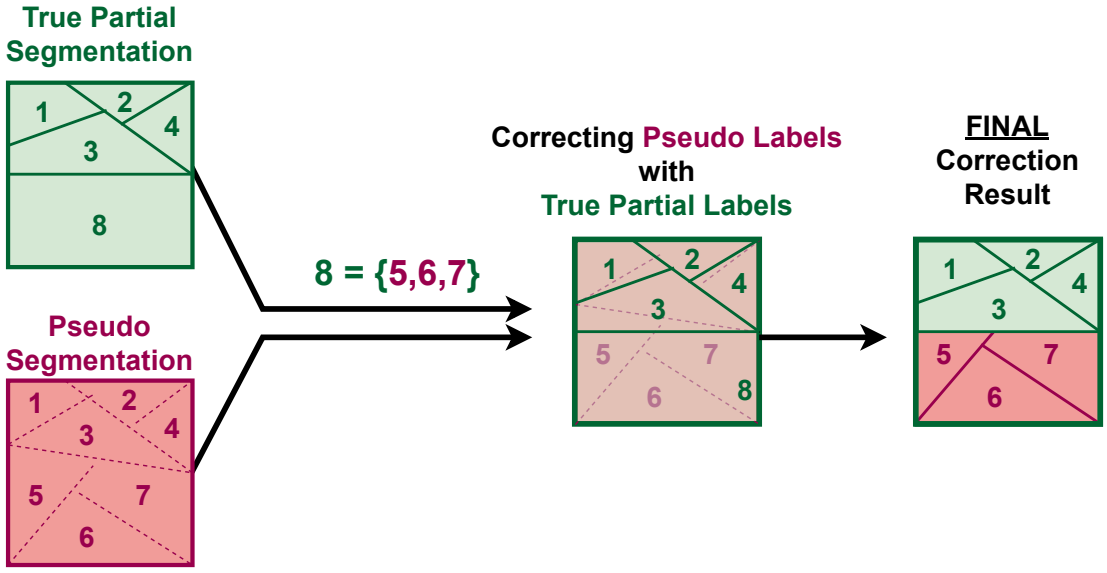


Figure 3: Example of pseudo-labels correction made with partial labels in the semi-supervised algorithm. 1,2,3,4,5,6,7,8 - are arbitrary segmentation classes. 5,6,7 are sub-labels of super-label 8. Pixels annotated as 5,6 and 7 that are outside the (known) region 8, are assigned their correct label. Pixels within region 8 that are initially annotated as one of {1,2,3,4} will be substituted for the most probable annotation from {5,6,7}.

5.2. Results

Our results show that supervised and semi-supervised models perform well when provided with 10% of fully labeled data (22 segmented MRI scans). However, their segmentation accuracy decreases when the number of annotations for sub-labels decreases. This can be observed, from the averaged Dice scores (twice the size of $Intersection/Union$ of the sets) shown for preterm and at-term neonatal scans in Fig. 4, where our model trained with $\hat{\mathcal{L}}_{\text{IUC}}$ loss generates more accurate segmentations than supervised and semi-supervised approaches trained with BCE loss. Our method achieves a dice score of 0.910 ± 0.03 for DGM with only 2%(4 scans) of the training samples having DGM annotated, while other models only reach accuracy of 0.709 ± 0.19 and 0.756 ± 0.17 for the case.

5.3. Hausdorff distance assessment

We computed the Hausdorff distance (HD) Taha and Hanbury (2015) for the dHCP dataset, with detailed results provided in the Appendix. While HD is known to be sensitive to outlier regions and boundary noise, which can lead to variability in performance, our method consistently demonstrates lower variance across most configurations. This stability, even when average HD values are close to those of the pseudo-labeling baseline, suggests that our model produces more reliable and consistent segmentations. Particularly in the 10%

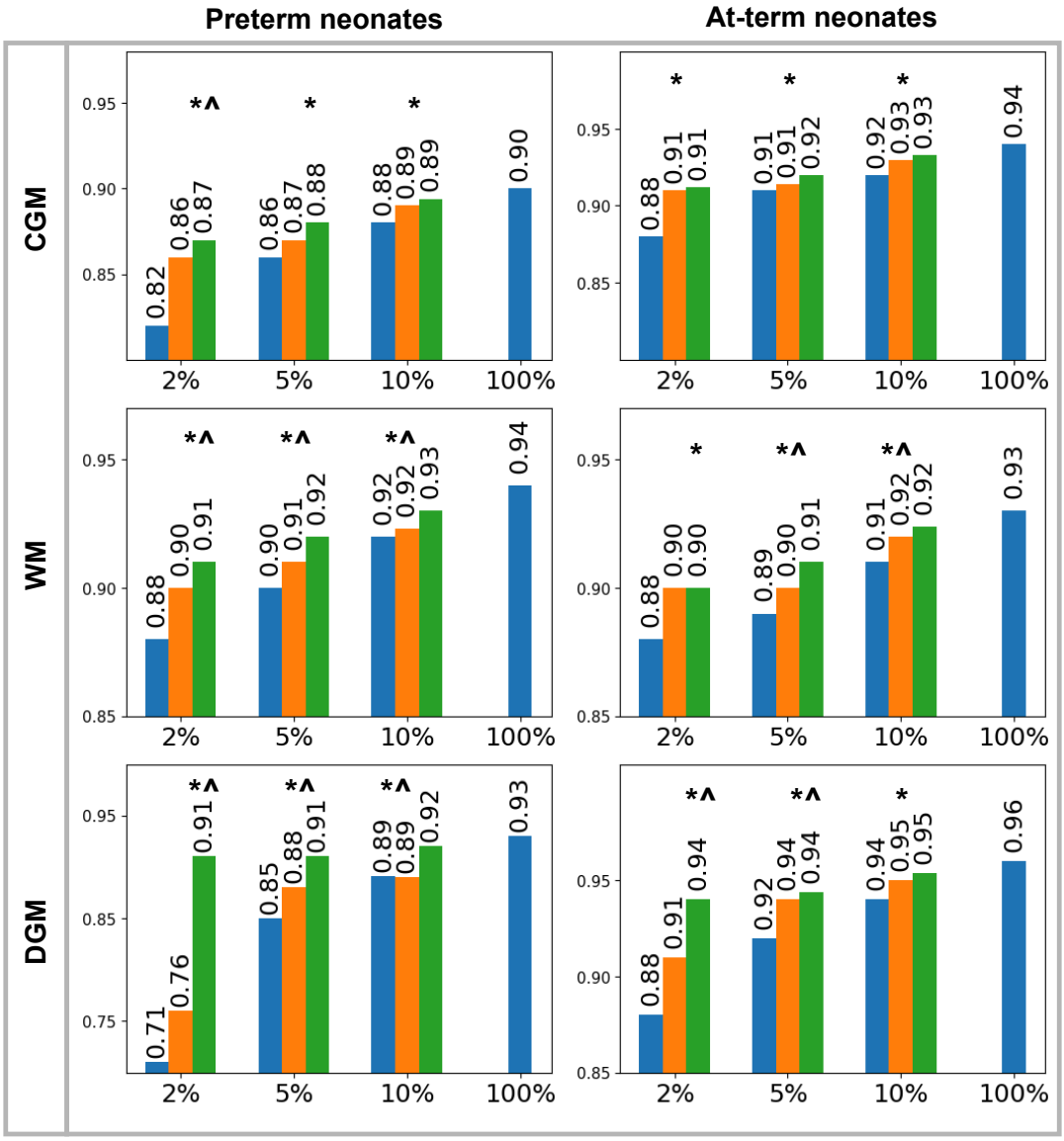


Figure 4: Average Dice scores on the dHCP dataset for preterm and at-term neonates using varying fractions $x = (2\%, 5\%, 10\%)$ of fully labeled training data alongside pseudo-labels for the Semi-Supervised model and partial labels for our model(green). Results show **superior performance of our method** on labels absent in the partial training data. The '*' and 'ˆ' signs mean our model's results are statistically significantly better ($p < 0.05$, ANOVA) than the supervised and semi-supervised models, accordingly.

Table 1: Average Dice scores for Deep Grey Matter (DGM), White Matter (WM), and Cortical Grey Matter (CGM) across models trained with different loss functions using 2% of fully labeled training data alongside pseudo-labels for the Semi-Supervised model and partial labels for our model. The table compares the effectiveness of each loss function in leveraging limited fully labeled data, highlighting the superior performance of the proposed Ignore Unobserved Child loss $\hat{\mathcal{L}}_{\text{IUC}}$ in low-label settings(2%, or 4 fully labeled scans), where noisy supervision can degrade performance. Scores are shown as mean±std.

Loss Function	PRETERM DGM	PRETERM WM	PRETERM CGM	AT-TERM DGM	AT-TERM WM	AT-TERM CGM
Supervised model						
$\hat{\mathcal{L}}_{\text{IUC}}$	0.71±0.19	0.88±0.02	0.82±0.06	0.88±0.08	0.88±0.02	0.88±0.02
Semi-Supervised model						
$\hat{\mathcal{L}}_{\text{IUC}}$	0.76±0.19	0.90±0.02	0.86±0.04	0.91±0.05	0.90±0.02	0.91±0.01
Our Model						
$\hat{\mathcal{L}}_{\text{IUC}}$	**0.91±0.03**	**0.91±0.01**	**0.87±0.03**	**0.94±0.01**	**0.90±0.02**	**0.91±0.01**
$\hat{\mathcal{L}}_{\text{WEIGHT}}$	0.65±0.04	0.87±0.01	0.81±0.04	0.70±0.01	0.84±0.02	0.87±0.01
$\hat{\mathcal{L}}_{\text{WNORM}}$	0.87±0.05	0.85±0.04	0.76±0.05	0.91±0.01	0.85±0.02	0.85±0.02
$\hat{\mathcal{L}}_{\text{HUC}}$	0.90±0.03	0.90±0.01	0.85±0.03	0.93±0.01	0.88±0.02	0.90±0.01

labeled setting for preterm neonates, our method shows clear improvements. Across other sub-regions, results are generally comparable, with a performance difference often exceeding 1σ, reinforcing the robustness of our approach in handling partial supervision.

5.4. Loss Functions

Table 1 highlights a key finding: the Ignore Unobserved Child loss $\hat{\mathcal{L}}_{\text{IUC}}$ consistently outperforms both the pseudo-labeling approach with knowledge priors (see Fig.4) and other loss functions incorporating pseudo-labels—especially in the low-label regime of just 2% fully labeled data (equivalent to only 4 scans). As the proportion of fully labeled data increases to 5%, performance across all loss functions converges as shown in Table 2. This suggests that in extremely low-label regimes, using a small amount of accurate ground truth can be as effective—or even preferable—compared to introducing additional supervision through potentially noisy pseudo-labels.

5.5. Architecture-agnostic methodology

Our approach is inherently free from strict architectural dependencies, making it adaptable to a wide range of segmentation models. To support this claim, we trained a SegFormer-B0 (Xie et al., 2021) architecture within our pipeline using IUC loss.

Table 3 shows that our method consistently outperforms both supervised and semi-supervised baselines in the low-label (2%) setting. For example, in the preterm group, our

Table 2: Average Dice scores for Deep Grey Matter (DGM), White Matter (WM), and Cortical Grey Matter (CGM) across models trained with different loss functions using 5% of fully labeled training data alongside pseudo-labels for the Semi-Supervised model and partial labels for our model. The table compares the effectiveness of each loss function in leveraging limited fully labeled data, highlighting that when 5% of the data is fully labeled, performance across all loss functions becomes comparable, suggesting that reliance on even modest amounts of accurate ground truth can match and **outperform** methods that rely on noisy pseudo-labels in low-label settings. Scores are shown as mean±std.

Loss Function	PRETERM			AT-TERM		
	DGM	WM	CGM	DGM	WM	CGM
Supervised model						
$\hat{\mathcal{L}}_{\text{IUC}}$	0.85±0.09	0.90±0.02	0.86±0.04	0.92±0.02	0.89±0.01	0.91±0.01
Semi-Supervised model						
$\hat{\mathcal{L}}_{\text{IUC}}$	0.88±0.08	0.91±0.01	0.87±0.03	0.94±0.01	0.90±0.01	0.91±0.01
Our Model						
$\hat{\mathcal{L}}_{\text{IUC}}$	0.91±0.03	0.92±0.01	0.88±0.02	0.94±0.01	0.91±0.01	0.92±0.01
$\hat{\mathcal{L}}_{\text{WEIGHT}}$	0.90±0.04	0.91±0.01	0.88±0.02	0.94±0.01	0.90±0.01	0.92±0.01
$\hat{\mathcal{L}}_{\text{WNORM}}$	0.91±0.02	0.91±0.01	0.87±0.03	0.94±0.01	0.90±0.01	0.91±0.01
$\hat{\mathcal{L}}_{\text{HUC}}$	0.92±0.03	0.92±0.01	0.88±0.02	0.95±0.01	0.91±0.01	0.92±0.01

model achieves a DGM Dice of 0.91 ± 0.03, compared to 0.82 ± 0.09 (supervised) and 0.82 ± 0.10 (semi-supervised). A similar trend is observed in the at-term group, where our model reaches 0.94 ± 0.01 versus 0.91 ± 0.03 and 0.91 ± 0.02, respectively. When 5% of the data is fully labeled, all models converge to similar performance levels, with our approach still achieving competitive results, confirming its robustness and architecture-agnostic applicability even in low-supervision regimes.

6. Discussion

Our research addresses a critical need for automated segmentation in neonatal brain MRI analysis. By developing a method that effectively utilizes partially labeled datasets, we offer a solution that improves segmentation accuracy without the need for extensive fully labeled training data.

In our study, we conducted comprehensive comparisons between our proposed model and commonly used supervised and semi-supervised methods. Our model consistently outperformed these approaches in terms of segmentation accuracy, especially in challenging scenarios such as very preterm neonatal scans. These results demonstrate superiority of our method and highlight its potential to significantly advance automated MRI segmentation in medical practice.

Table 3: Average Dice scores for Deep Grey Matter (DGM), White Matter (WM), and Cortical Grey Matter (CGM) across models trained with SegFormer-B0 architecture and Ignore Unobserved Child loss $\hat{\mathcal{L}}_{\text{IUC}}$ using varying fractions (2%, 5%) of fully labeled training data alongside pseudo-labels for the Semi-Supervised model and partial labels for our model. The table presents performance gains of our method comparable to those achieved with U-Net in low-label settings(2%, or 4 fully labeled scans), where noisy supervision can degrade performance. When 5% of the data is fully labeled, performance across all models becomes comparable, suggesting that reliance on even modest amounts of accurate ground truth can match and **outperform** methods that rely on noisy pseudo-labels in low-label settings. Scores are shown as mean±std. **Bold** values indicate statistically significant improvement over the supervised model, and *italic* values indicate improvement over the semi-supervised model ($p < 0.05$, ANOVA).

PRETERM			AT-TERM		
DGM	WM	CGM	DGM	WM	CGM
Supervised model (Trained on 2% Fully Labeled)					
0.82±0.09	0.89±0.01	0.81±0.04	0.91±0.03	0.86±0.03	0.87±0.01
Semi-Supervised model (Trained on 2% Fully Labeled + 98% Pseudo Labeled Data)					
0.82±0.10	0.89±0.01	0.83±0.04	0.91±0.02	0.86±0.03	0.88±0.01
Our Model (Trained on 2% Fully Labeled + 98% Partially Labeled Data)					
0.91±0.03	***0.90±0.01***	**0.84±0.03**	***0.94±0.01***	**0.87±0.02**	**0.88±0.01**
Supervised model (Trained on 5% Fully Labeled)					
0.89±0.04	0.90±0.01	0.84±0.03	0.93±0.01	0.87±0.02	0.88±0.01
Semi-Supervised model (Trained on 5% Fully Labeled + 95% Pseudo Labeled Data)					
0.91±0.03	0.91±0.01	0.85±0.03	0.94±0.01	0.87±0.02	0.89±0.01
Our Model (Trained on 5% Fully Labeled + 95% Partially Labeled Data)					
0.91±0.03	0.90±0.01	0.85±0.03	**0.94±0.01**	0.87±0.02	0.88±0.01

This approach not only enhances the efficiency of MRI analysis, but also has broader implications beyond neonatal brain segmentation, extending to various medical imaging tasks where multi-label partially annotated data is common.

Limitations While our method demonstrates substantial benefits from combining larger-scale labels with a limited number of detailed segmentations, it does require the presence of at least some segmentation data to fully leverage its advantages. As such, it is not designed for entirely unsupervised scenarios where no segmentation annotations are available.

The approach is particularly well-suited for applications where detailed annotation is inherently challenging or resource-intensive—such as in the case of preterm neonates, where anatomical variability and low image contrast complicate traditional segmentation methods. However, in domains where high-resolution images and clearly defined structures are readily available, such as adult brain MRI, existing methods trained on a small number of detailed annotations may already perform sufficiently well, potentially limiting the added value of our approach in those settings.

Acknowledgments

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute https://vectorinstitute.ai/partnerships/current-partners/. MB is a CIFAR CCAI Chair.

References

John Ashburner. Spm: A history. *NeuroImage*, 62(2):791–800, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2011.10.025. URL https://www.sciencedirect.com/science/article/pii/S1053811911011888. 20 YEARS OF fMRI.

Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4764–4772, 2022.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels, 2021.

Andrew Joseph Davies. Semantic segmentation of aerial imagery using u-net in python, 2022. URL https://towardsdatascience.com/semantic-segmentation-of-aerial-imagery-using-u-net-in-python-552705238514.

Ahmed E. Fetit, John Cupitt, Turkay Kart, and Daniel Rueckert. Training deep segmentation networks on texture-encoded input: application to neuroimaging of the developing neonatal brain. In Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal, editors, *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 230–240. PMLR, 06–08 Jul 2020. URL https://proceedings.mlr.press/v121/fetit20b.html.

Bruce Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2012.01.021. URL https://www.sciencedirect.com/science/article/pii/S1053811912000389. 20 YEARS OF fMRI.

Emer J. Hughes, Tobias Winchman, Francesco Padormo, Rui Teixeira, Julia Wurie, Maryanne Sharma, Matthew Fox, Jana Hutter, Lucilio Cordero-Grande, Anthony N. Price, Joanna Allsop, Jose Bueno-Conde, Nora Tusor, Tomoki Arichi, A. D. Edwards, Mary A. Rutherford, Serena J. Counsell, and Joseph V. Hajnal. A dedicated neonatal brain imaging system. *Magnetic Resonance in Medicine*, 78(2):794–804, 2017. doi: https://doi.org/10.1002/mrm.26462. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.26462.

Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. Fsl. *NeuroImage*, 62(2):782–790, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2011.09.015. URL https://www.sciencedirect.com/science/article/pii/S1053811911010603. 20 YEARS OF fMRI.

Suhyeon Jo, DongHyeok Shin, Byeonghu Na, JoonHo Jang, and Il-Chul Moon. Hierarchical multi-label classification with partial labels and unknown hierarchy. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 1025–1034, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3614912. URL https://doi.org/10.1145/3583780.3614912.

Jeremy Jordan. An overview of semantic image segmentation, 2018. URL https://www.jeremyjordan.me/semantic-segmentation/.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

Michael Lempart, Martin P. Nilsson, Jonas Scherman, Christian Jamtheim Gustafsson, Mikael Nilsson, Sara Alkner, Jens Engleson, Gabriel Adrian, Per Munck af Rosenschöld, and Lars E. Olsson. Pelvic u-net: multi-label semantic segmentation of pelvic organs at risk for radiation therapy anal cancer patients using a deeply supervised shuffle attention convolutional neural network. *Radiation Oncology*, 17(1), December 2022. ISSN 1748-717X. doi: 10.1186/s13014-022-02088-1.

Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1246–1257, June 2022.

Meichen Liu, Xin Yan, Chenhui Wang, and Kejun Wang. Segmentation mask-guided person image generation. *Applied Intelligence*, 51(2):1161–1176, feb 2021. ISSN 0924-669X. doi: 10.1007/s10489-020-01907-w. URL https://doi.org/10.1007/s10489-020-01907-w.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Marcel Prastawa, John H. Gilmore, Weili Lin, and Guido Gerig. Automatic segmentation of mr images of the developing newborn brain. *Medical Image Analysis*, 9(5):457–466,

2005. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2005.05.007. URL https://www.sciencedirect.com/science/article/pii/S1361841505000630. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2004.

Simon Reiss, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9532–9542, June 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Abdel Aziz Taha and Allan Hanbury. An efficient algorithm for calculating the exact hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37 (11):2153–2163, 2015. doi: 10.1109/TPAMI.2015.2408351.

Romain Valabrègue, F Girka, Alexandre Pron, François Rousseau, Guillaume Auzias, du Cerveau, and Context. Comprehensive analysis of synthetic learning applied to neonatal brain mri segmentation. *Human Brain Mapping*, 45, 2023. URL https://api.semanticscholar.org/CorpusID:261681857.

Ao Wang, Hui Chen, Zijia Lin, Zixuan Ding, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Guiguang Ding. Hierarchical prompt learning using clip for multi-label classification with single positive labels. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5594–5604, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Han Zheng, Lanfen Lin, Hongjie Hu, Qiaowei Zhang, Qingqing Chen, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Ruofeng Tong, and Jian Wu. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 148–156. Springer, 2019.

Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, page 423–440, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_25. URL https://doi.org/10.1007/978-3-031-20053-3_25.

Lilla Zöllei, Juan Eugenio Iglesias, Yangming Ou, P. Ellen Grant, and Bruce Fischl. Infant freesurfer: An automated segmentation and surface extraction pipeline for t1-weighted neuroimaging data of infants 0–2 years. *NeuroImage*, 218:116946, 2020.

ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.116946. URL https://www.sciencedirect.com/science/article/pii/S1053811920304328.

Appendix A.

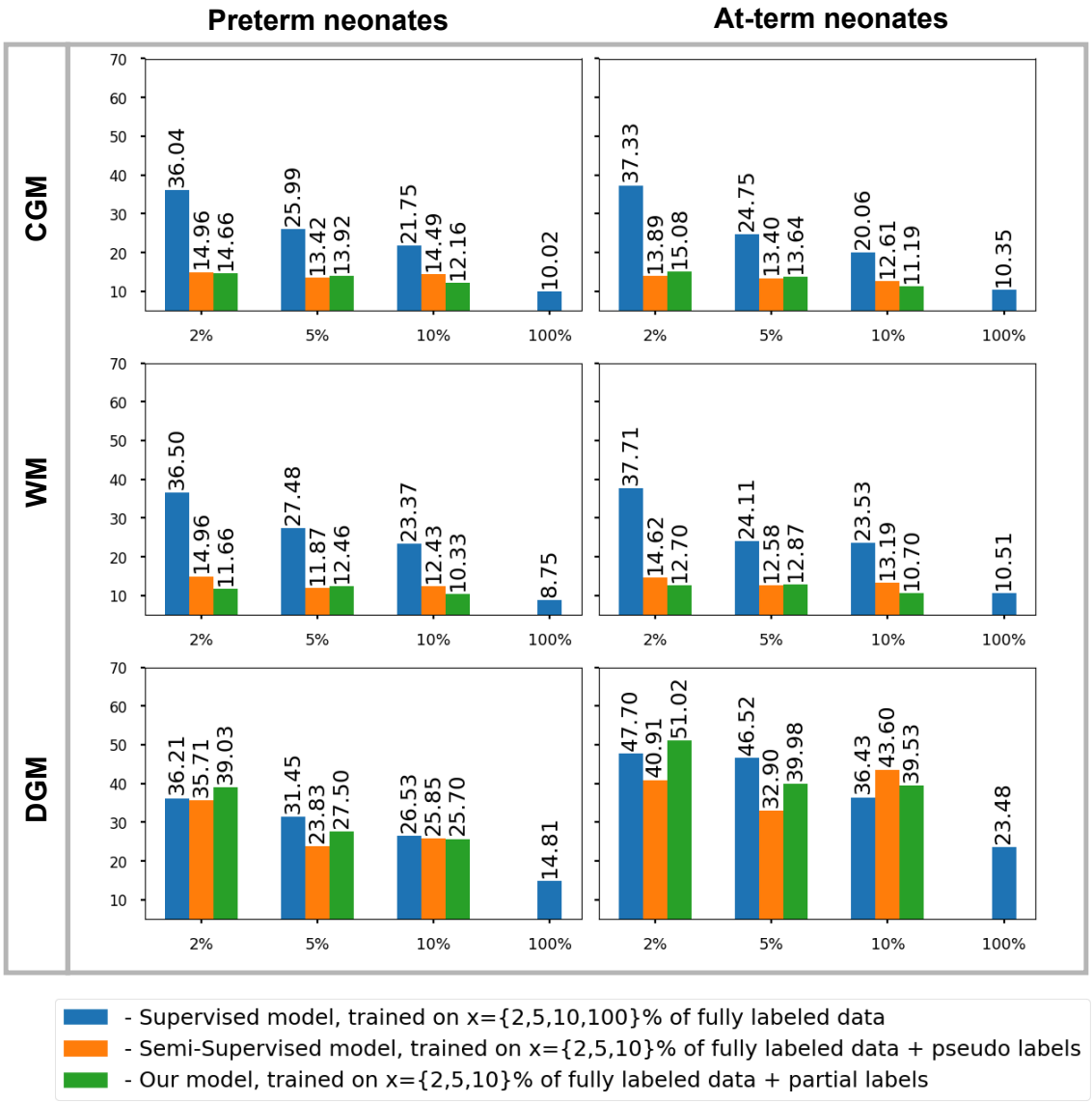
Hausdorff distance results with IUC loss

	Preterm			At-Term		
	Cortical gray matter	White matter	Deep Gray Matter	Cortical gray matter	White matter	Deep Gray Matter
2%	36.044 ±6.54	36.501 ±6.24	36.211 ±8.96	37.331 ±6.28	37.707 ±5.86	47.704 ±10.29
	14.957 ±3.94	14.961 ±2.46	35.709 ±10.0	13.889 ±3.94	14.615 ±3.6	40.911 ±8.67
	14.665 ±1.8	**11.66 ±2.26**	39.034 ±9.54	15.078 ±2.94	**12.695 ±3.09**	51.025 ±8.25
	Preterm			At-Term		
5%	25.989 ±9.08	27.482 ±9.14	31.45 ±10.54	24.75 ±10.68	24.109 ±10.83	46.516 ±12.56
	13.424 ±1.49	11.872 ±1.63	23.826 ±9.27	13.397 ±2.24	12.582 ±1.97	32.901 ±10.04
	13.916 ±1.88	12.463 ±1.44	27.5 ±8.38	13.64 ±1.4	12.868 ±1.48	39.983 ±8.06
	Preterm			At-Term		
10%	21.754 ±11.01	23.374 ±10.17	26.53 ±8.33	20.057 ±10.06	23.529 ±9.41	36.431 ±8.77
	14.487 ±4.41	12.431 ±4.23	25.852 ±8.52	12.611 ±2.54	13.194 ±3.17	43.603 ±8.99
	12.157 ±2.49	**10.326 ±2.19**	**25.698 ±7.53**	**11.185 ±2.59**	**10.695 ±2.04**	39.528 ±9.39
100%	10.016 ±1.66	8.747 ±0.76	14.814 ±4.85	10.353 ±2.13	10.513 ±1.72	23.485 ±12.85

- Supervised model, trained on x={2,5,10,100}% of fully labeled data
- Semi-Supervised model, trained on x={2,5,10}% of fully labeled data + pseudo labels
- Our model, trained on x={2,5,10}% of fully labeled data + partial labels

Average Hausdorff Distance values for models trained with IUC loss for preterm and at-term neonates using varying fractions $x = (2\%, 5\%, 10\%)$ of fully labeled training data alongside pseudo labels for the Semi-Supervised model and partial labels for our model(green).

Appendix B.



Bar plot for average Hausdorff Distance values for models trained with IUC loss for preterm and at-term neonates using varying fractions $x = (2\%, 5\%, 10\%)$ of fully labeled training data alongside pseudo labels for the Semi-Supervised model and partial labels for our model(green).