

Uncertainty estimates as data selection criteria to boost omni-supervised learning

Lorenzo Venturini¹, Aris T. Papageorgiou², J. Alison Noble¹, and Ana I.L. Namburete¹

¹ Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom

² Nuffield Department of Women’s and Reproductive Health, University of Oxford, Oxford, United Kingdom

Abstract. For many medical applications, large quantities of imaging data are routinely obtained but it can be difficult and time-consuming to obtain high-quality labels for that data. We propose a novel uncertainty-based method to improve the performance of segmentation networks when limited manual labels are available in a large dataset. We estimate segmentation uncertainty on unlabeled data using test-time augmentation and test-time dropout. We then use uncertainty metrics to select unlabeled samples for further training in a semi-supervised learning framework. Compared to random data selection, our method gives a significant boost in Dice coefficient for semi-supervised volume segmentation on the EADC-ADNI/HARP MRI dataset and the large-scale INTERGROWTH-21st ultrasound dataset. Our results show a greater performance boost on the ultrasound dataset, suggesting that our method is most useful with data of lower or more variable quality.

Keywords: Uncertainty, Omni-supervised learning, Boosting

1 Introduction

A major challenge in supervised medical image segmentation is the scarcity of accurately labeled data [13]. While imaging is routinely used in clinical settings, it is much harder to obtain accurate segmentation labels for imaging data. Producing accurate manual labels, especially for 3D segmentation tasks, is time-consuming and often not part of a standard clinical protocol.

Machine learning methods such as convolutional neural networks (CNNs) typically rely on large amounts of labeled data to achieve optimal performance, but generating reliable manual labels for medical datasets requires significant time investment from trained clinicians, which is often not available.

Several approaches have been proposed to reduce the time investment required to generate labels, such as weak supervision with bounding boxes [11] or annotated landmarks [1]. There have also been attempts to use additional unlabeled data to improve a network’s performance. Active learning methods [19] can prompt a human labeler to only provide additional manual labels on examples that would lead to the greatest performance improvements. Adversarial

networks have also been proposed to improve performance with limited training labels [20].

Another method that has been proposed to exploit the presence of unlabeled data to improve a neural network’s performance is omni-supervised learning [16]. In general, training a neural network using its own predictions does not improve its performance [11], but combining the predictions of multiple independent learners leads to a prediction superior to any of them [5,7]. Similarly, aggregating predictions generated from different transformations (such as reflections and rotations) of the data can also lead to improvements [12,16]. In other words, “a set of weak learners can create a strong learner” [17]. These two principles, of *model diversity* and *data diversity*, drive the concept of omni-supervised learning.

Omni-supervised learning aggregates network-generated segmentations of the unlabeled dataset using both types of diversity and uses them additional labels for further training. Huang et al [8] used this to improve localisation in 3D neurosonography images. In that work, a subset of the predictions generated on unlabeled data are used for the next iteration of training. Selecting an appropriate subset of predictions, however, has been ignored in the literature. Previous work using CNNs for omni-supervised learning has either selected a subset of the unlabeled dataset at random [8] or used weak heuristics [16], such as ensuring that the number of labeled pixels is similar to the average in the training data. In this work, we explore different selection criteria and evaluate their performance.

1.1 Uncertainty estimation

CNNs for binary segmentation typically use sigmoid activation functions at their output, so each voxel is given a “soft” classification ranging from 0 (confident background) to 1 (confident foreground). The network thus expresses some uncertainty in its segmentation, but CNNs also often confidently make incorrect predictions [4], especially when they are presented with noisy or ambiguous data, or when the data presented to them differs in appearance from the training data.

A distinction can be made between the two sources of error: (1) error caused by image noise, ambiguity, and inconsistent labeling in the training set, and (2) error caused by the classifier being shown data that appears different to that in the training set. The first type, known as *aleatoric uncertainty* [10], is due to ambiguity and noise in the data itself. Since this is a measure of genuine uncertainty in the data, this type of uncertainty is unlikely to reduce with increased training data. On the other hand, the second source of error, known as *epistemic uncertainty* [10], is due to the model parameters and overfitting to data not represented in the training set. The amount of epistemic uncertainty can be used to give a measure of how different any given example is from the training data.

Gal and Ghahramani [6] have proposed a method to estimate the uncertainty present in the model, using *dropout uncertainty*. Their method applies dropout at test time and measures differences in output to give a measure of the model’s uncertainty. Wang et al. [18] proposed using test-time augmentation to

improve estimates of aleatoric uncertainty, using a similar method: measuring differences in output across augmented versions of the same data. The principle behind these differing approaches comes from different sources of uncertainty. Epistemic uncertainty, introduced by the model’s parameters, is estimated by making changes to the model. Aleatoric uncertainty is introduced by the data, so it can be estimated by varying the data.

Omni-supervised learning requires data diversity and model diversity, which are also methods to measure aleatoric and epistemic uncertainty, respectively. In this paper we show that these two measures can form the selection criteria of data for omni-supervised models, without adding complexity.

We propose a scheme to incorporate these metrics of uncertainty in an omni-supervised framework to select data to use for subsequent training iterations. We harness data and model diversity in omni-supervised learning to obtain uncertainty estimates at little additional computational cost. We then experiment with data selection for further training based on these metrics, and compare the segmentation performance of different data-selection methods with random selection. Our novel data-selection scheme uses information from the unlabeled data to improve segmentation performance with no additional supervision.

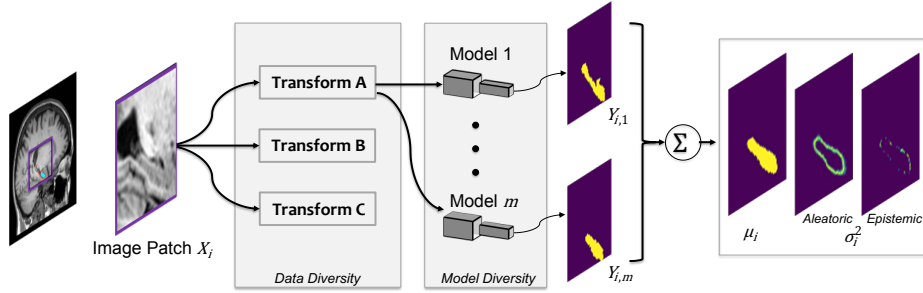


Fig. 1. Data diversity consisting of transformations and model diversity using m models are used to generate independent predictions of the same sample X_i , which can be aggregated to generate a stronger prediction.

2 Methods

2.1 Uncertainty estimation

To measure epistemic uncertainty on 3D data we implemented a minor variation on a conventional 3D U-Net [3] adding dropout to obtain uncertainty estimates. After each max-pool layer or before each upsampling layer, some of the previous layer’s weights are dropped out ($p = 0.5$).

Using test-time dropout, we generated N_p independent predictions for each unlabeled volume. The uncertainty in the segmentation of each voxel is then

given by the variance in the voxel’s predicted label across different segmentations³. A global measure of the epistemic uncertainty in a volume can be obtained simply by summing the uncertainty of all pixels

$$\text{uncertainty} = \sum_i \sigma_n^2(x_i)$$

where $\sigma_n^2(x_i)$ is the variance of the softmax score of each pixel x_i over n samples.

While epistemic uncertainty arises from the model, and can be measured by varying the model at test time, aleatoric uncertainty arises from the data, and can be measured by varying the data at test time. The augmentation methods used are described in section 2.2.

While higher uncertainty across both metrics, in general, appears to point to a lower-quality segmentation, it is also important to increase the diversity of the training dataset and reduce the differences between the training data and the unlabeled data. We hypothesise, therefore, that selecting volumes segmented with a lower level of epistemic uncertainty will not lead to an improvement in segmentation performance in the next round of training.

The same is not true of *aleatoric* uncertainty, which is linked to the data, rather than the model: higher aleatoric uncertainty is likely to link to a lower signal-to-background ratio, or more challenging data. Therefore, we expect that selecting volumes with lower aleatoric uncertainty will lead to better results.

2.2 Experiments

We validated our hypothesis on models built on two datasets, the EADC-ADNI/HARP MRI dataset and the large-scale INTERGROWTH-21st ultrasound dataset.

MRI dataset: The ADNI initiative is an ongoing multisite longitudinal study that acquires structural MRI brain volumes from cognitively normal aging adults (CN), as well as those with mild cognitive impairment (MCI) and Alzheimer’s disease (AD) [9]. 63 sites participate in acquisition, using different scanners and acquisition methods. The EADC/HARP dataset consists of a subset of 135 volumes selected from the ADNI dataset, with expert 3D manual segmentations of the hippocampus. We also used 680 additional unlabeled volumes from the ADNI dataset, each from different subjects, removing any volumes already in the HARP dataset. All volumes were brain-extracted, linearly registered to MNI152 image space [2] and normalized by dividing each pixel by the 99th percentile value. $64 \times 64 \times 64$ patches were extracted around the hippocampus in each hemisphere, leaving 270 labeled patches and 1360 unlabeled patches. Of the 135 labeled volumes, 80 for training, 20 for validation, and 35 for testing. Five-fold cross-validation was used.

³ As most voxels distant from anatomical boundaries are consistently segmented, the uncertainty (or segmentation variance) of most voxels is 0.

Ultrasound dataset: We used data from the INTERGROWTH-21st study [15], a longitudinal study which includes 3D ultrasound volumes acquired from optimally healthy pregnant women. For this study, we used 948 neurosonography volumes in the range of 14-25 gestational weeks. The volumes were registered to a common coordinate system using similarity transforms [14]. The cerebellum was manually segmented by the same annotator on 146 volumes: 86 for training, 20 for validation, and 40 for testing. Five-fold cross-validation was used. The remaining volumes were kept unlabeled.

We trained a model for each dataset using the labeled volumes in the training set, and used the model to generate segmentations. The aleatoric and epistemic uncertainties of each segmentation were estimated using test-time augmentation and test-time dropout.

An aggregated segmentation was produced for each volume by taking the median predicted value for each pixel. N_{unl} volumes were then selected for the second stage of omni-supervised training: these volumes (with corresponding annotations) are added to the training set and the model is retrained using the expanded training set. We experimented with different selection criteria for the automatically segmented volumes:

1. Randomly select the volumes, as a control (similar to current methods [8])
2. Select the volumes with the lowest epistemic uncertainty
3. Select the volumes with the lowest aleatoric uncertainty.
4. Select the volumes with the lowest sum of epistemic and aleatoric uncertainty.

We explored the effect of different amounts of initial labeled data. More training data would be expected to reduce epistemic uncertainty, as the training set more closely resembles the distribution of possible images and the network’s overall performance improves, but should not affect aleatoric uncertainty estimates.

Implementation details: The augmentation used consists of simple similarity transforms. Since the images are aligned to a common reference space (in both datasets), the only axis they can be reflected across to maintain alignment is the midsagittal plane (MSP), the median plane that runs vertically between the brain’s hemispheres. Small random **translations** (up to ± 5 voxels along each axis), **rotations** (up to $\pm 10^\circ$ around each axis), and **scaling** (up to $\pm 10\%$ linear zoom) were also used. Nearest-neighbour interpolation was used for computational efficiency. The volumes were **reflected** with 50% probability, while the degree of all other augmentations was sampled from a uniform distribution. Once a prediction had been obtained on an augmented volume, the inverse transform T_x^{-1} was applied to return to a common reference space. Aleatoric uncertainty was estimated using the variance of the resulting predictions, after test-time augmentation. The variance was measured across N_p predictions per volume. This augmentation scheme was used for all experiments.

We selected $N_{unl} = 200$ for the MRI dataset and $N_{unl} = 150$ volumes for the ultrasound dataset, to be similar in size to the labeled training dataset in line

with previous work on omni-supervised learning [8]. We also selected $N_p = 40$, to allow an accurate estimate of variance without using excessive computational resources.

3 Results

3.1 Uncertainty quantification

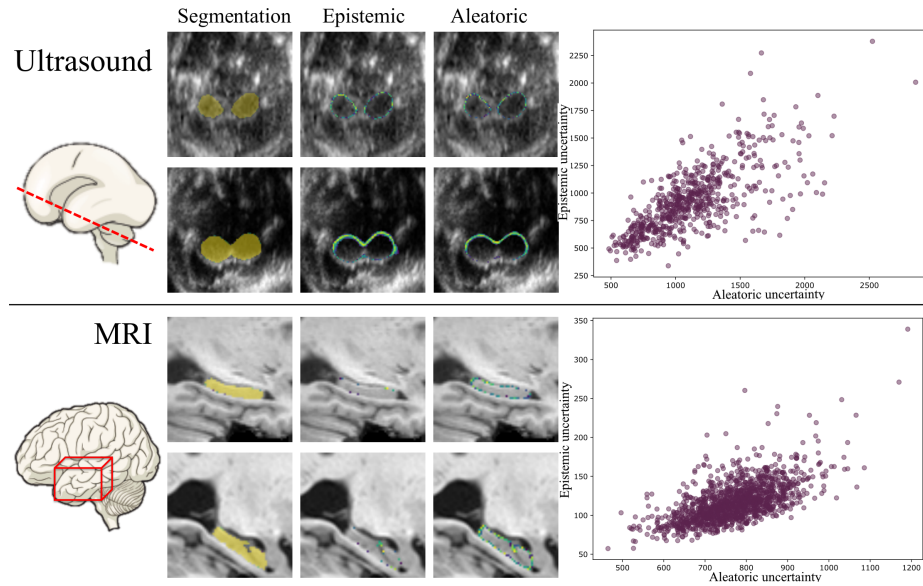


Fig. 2. Visualisation of aleatoric and epistemic uncertainty estimates in two example images per dataset, and scatterplots (with transparency for overlapping points) showing uncertainty across both datasets.

Figure 2 shows global estimates of aleatoric and epistemic uncertainty of every volume in the unlabeled dataset. Due to less well-defined edges in ultrasound and higher variability in image quality, the ultrasound dataset shows higher uncertainty in both metrics as seen by the higher values on the axes. The bottom ultrasound example shows a strong shadow obscuring part of the cerebellum, and uncertainty is accordingly higher in that area.

The plots in Figure 2 show a correlation between the estimates of aleatoric and epistemic uncertainty ($r = 0.72$ in the ultrasound dataset, and $r = 0.69$ in the MRI dataset). We believe this is partially driven by boundary effects: classification is more difficult at boundary voxels, so structures with a higher surface area are likely to show higher uncertainty in both cases.

3.2 Selection methods

| | Ultrasound | | MRI | |
|------------------|-------------------------------------|-----------------------------------|-------------------------------------|-----------------------------------|
| Volume selection | Dice coeff. | Hausdorff (mm) | Dice coeff. | Hausdorff (mm) |
| Fully supervised | 0.673 ± 0.046 | 12.25 ± 3.57 | 0.848 ± 0.015 | 3.97 ± 0.25 |
| Random | 0.700 ± 0.023 | 11.44 ± 1.75 | 0.849 ± 0.015 | 3.73 ± 0.18 |
| Epistemic | 0.689 ± 0.040 | 10.01 ± 1.48 | 0.847 ± 0.015 | 4.15 ± 0.32 |
| Aleatoric | 0.727 ± 0.014 | 8.54 ± 1.60 | 0.851 ± 0.015 | 3.69 ± 0.25 |
| Aleat. + epist. | 0.697 ± 0.015 | 11.10 ± 1.51 | 0.848 ± 0.015 | 4.20 ± 0.35 |
| IOV | 0.764 ± 0.060 | 3.96 ± 0.99 | N/A | N/A |

Table 1. The segmentation performance of retraining a 3D CNN using different selection methods for the additional labeled data. The “IOV” row indicates the intra-observer variability of manual segmentations obtained by the same individual.

Table 1 shows the performance of each data selection method. In ultrasound, random volume selection leads to an improved Dice coefficient ($p < 0.02$ in a 1-tailed t -test) in the validation set over a fully-supervised network, but does not lead to a significant improvement in Hausdorff distance. Selecting the volumes with the lowest epistemic uncertainty leads to no significant improvement ($p = 0.08$) in Dice coefficient or Hausdorff distance. Selecting volumes based on the lowest aleatoric uncertainty, on the other hand, led to significant improvements across both metrics ($p < 0.01$). Selecting volumes based on the average of aleatoric and epistemic uncertainty seems to yield accuracy roughly in the middle of the two methods ($p < 0.03$). This does not represent an improvement over aleatoric selection only.

In the MRI dataset, effects are smaller: the effect size is not significant ($p = 0.1$) for random volume selection, but remains significant ($p < 0.01$) for aleatoric selection. The same pattern as with the ultrasound dataset emerges, with larger performance improvement with aleatoric selection over random selection.

One possible explanation for the small effect size is that the labeled data alone is sufficient to generalize to the test set, and the network is already near the maximum achievable performance on this dataset. Figure 3a shows that reducing the number of labeled training examples to 10 or 20 (from the original 180) seems to have only a limited impact on segmentation performance.

4 Discussion

We observe an increase in Dice coefficient using omni-supervised learning over a fully-supervised implementation. Selecting the volumes with the lowest epistemic uncertainty, measured by test-time dropout, seems to lead to worse performance than random selection, especially in Hausdorff distance. The volumes with the lowest epistemic uncertainty are those that most closely resemble the training

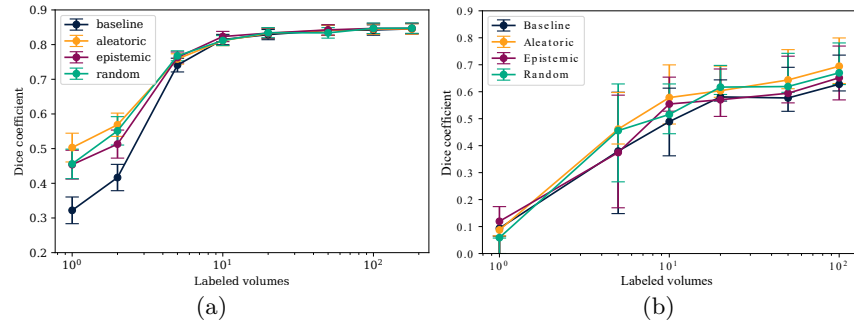


Fig. 3. Change in Dice coefficient with a changing number of labeled training examples for the (a) MRI dataset, and (b) the ultrasound dataset.

data, and we believe that adding these volumes to the training set does not improve the generalisation of the network. Instead, it introduces a bias in the training data towards the segmentations it has generated in the previous round. Further evidence of this is seen in the results selecting samples with the average of aleatoric and epistemic uncertainty: it is lower than with aleatoric uncertainty alone, suggesting that including epistemic uncertainty in this selection does not add value.

On the other hand, using aleatoric uncertainty leads to a significant improvement in performance over a random selection, especially on the ultrasound dataset. Aleatoric uncertainty is used as an estimate of the amount of uncertainty inherent in the data itself. In principle, it is unrelated to how different the data appears from training and only a measure of the quality of the data itself. Lower aleatoric uncertainty, therefore, can be thought of as correlating to clearer data and possibly better segmentation performance, without compromising the generalisability of the segmentation method.

The performance improvement is significantly larger in the ultrasound dataset. Direct comparisons are difficult since it's a different structure, but we believe that this is due to the lower signal-to-background ratio of ultrasound imaging, and the presence of artifacts such as shadows, in some volumes. The anatomical boundaries of the cerebellum in the volumes we examined were also ambiguous, which is itself a source of aleatoric uncertainty. We believe that the greater variation in image quality in ultrasound imaging accounts for the increased performance gain from using aleatoric uncertainty-based selection. This implies that a scheme similar to the one produced here may lead to a larger performance boost in datasets with data of lower, or more variable, quality.

Results on MRI nonetheless show a small, but significant increase in Dice coefficient. We believe this is partly because the fully-supervised network is already close to the highest performance a CNN-based method could achieve: Figure 3 shows only a modest performance reduction from a 10-fold reduction in training examples. With fewer training examples, the performance boost is larger on two medical datasets with the same segmentation CNN.

5 Conclusion

We demonstrate a novel uncertainty-based data selection scheme for omni-supervised learning, demonstrating its effectiveness for different segmentation tasks with different imaging modalities. We found a significant improvement for both segmentation problems which increases with fewer training labels, and found a larger performance boost in ultrasound. Our results suggest that our method is of most use in datasets with less labeled data, and with images of lower or more variable quality.

Acknowledgment

We would like to thank Nicola Dinsdale for her help with data preparation and analysis of the MRI dataset. This work is supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) [grant number EP/L016052/1]. A. T. Papageorgiou is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. A. Namburete is grateful for support from the UK Royal Academy of Engineering under the Engineering for Development Research Fellowships scheme. J. A. Noble acknowledges the National Institutes of Health (NIH) through the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (U01 AA014809-14). We thank the INTERGROWTH-21st Consortium for permission to use 3D ultrasound volumes of the fetal brain.

References

1. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2016). https://doi.org/10.1007/978-3-319-46478-7_34
2. Bocchetta, M., Boccardi, M., Ganzola, R., Apostolova, L.G., Preboske, G., Wolf, D., Ferrari, C., Pasqualetti, P., Robitaille, N., Duchesne, S., Jack, C.R., Frisoni, G.B., Bartzokis, G., Decarli, C., Detolledo-Morrell, L., Fellgiebel, A., Firbank, M., Gerritsen, L., Henneman, W., Killiany, R.J., Malykhin, N., Pruessner, J.C., Soininen, H., Wang, L.: Harmonized benchmark labels of the hippocampus on magnetic resonance: The EADC-ADNI project. *Alzheimer's and Dementia* (2015). <https://doi.org/10.1016/j.jalz.2013.12.019>
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–432. Springer (2016)
4. Denker, J.S., LeCun, Y.: Transforming Neural-Net Output Levels to Probability Distributions. *Advances in Neural Information Processing Systems* 3 (1991)
5. Dietterich, T.G.: Ensemble methods in machine learning. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2000)
6. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (jun 2015), <http://arxiv.org/abs/1506.02142>
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network (mar 2015), <http://arxiv.org/abs/1503.02531>
8. Huang, R., Noble, J.A., Namburete, A.I.L.: Omni-Supervised Learning: Scaling Up to Large Unlabelled Medical Datasets. pp. 572–580. Springer, Cham (sep 2018). https://doi.org/10.1007/978-3-030-00928-1_65, http://link.springer.com/10.1007/978-3-030-00928-1_65
9. Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W.: The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods (2008). <https://doi.org/10.1002/jmri.21049>
10. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? (2017), <http://papers.nips.cc/paper/7141-what-uncertainties-do-we-need>
11. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does It: Weakly supervised instance and semantic segmentation. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (2017). <https://doi.org/10.1109/CVPR.2017.181>
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems* (2012)

13. Makropoulos, A., Counsell, S.J., Rueckert, D.: A review on automatic fetal and neonatal brain MRI segmentation. *NeuroImage* (2017). <https://doi.org/10.1016/j.neuroimage.2017.06.074>
14. Namburete, A.I., Xie, W., Yaqub, M., Zisserman, A., Noble, J.A.: Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning. *Medical Image Analysis* **46**, 1–14 (may 2018). <https://doi.org/10.1016/J.MEDIA.2018.02.006>, <https://www.sciencedirect.com/science/article/pii/S1361841518300306><http://www.ncbi.nlm.nih.gov/pubmed/29499436>
15. Papageorgiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., Noble, J.A., Pang, R., Victora, C.G., Barros, F.C., Carvalho, M., Salomon, L.J., Bhutta, Z.A., Kennedy, S.H., Villar, J.: International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet* **384**(9946), 869–879 (2014). [https://doi.org/10.1016/S0140-6736\(14\)61490-2](https://doi.org/10.1016/S0140-6736(14)61490-2), <http://linkinghub.elsevier.com/retrieve/pii/S0140673614614902>
16. Radosavovic, I., Dollar, P., Girshick, R., Gkioxari, G., He, K.: Data Distillation: Towards Omni-Supervised Learning. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018). <https://doi.org/10.1109/CVPR.2018.00433>
17. Schapire, R.E.: The Strength of Weak Learnability. *Machine Learning* (1990). <https://doi.org/10.1023/A:1022648800760>
18. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (apr 2019). <https://doi.org/10.1016/J.NEUCOM.2019.01.103>, <https://www.sciencedirect.com/science/article/pii/S0925231219301961>
19. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) *MICCAI 2017*, pp. 399–407. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_46, https://doi.org/10.1007/978-3-319-66179-7_{_}46
20. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep Adversarial Networks for Biomedical Image Segmentation Utilizing Unannotated Images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) *MICCAI 2017*, pp. 408–416. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47, https://doi.org/10.1007/978-3-319-66179-7_{_}47