Semi-supervised Learning for Fetal Brain MRI Quality Assessment with ROI Consistency

Junshen Xu[1(✉)], Sayeri Lala[1], Borjan Gagoski[2], Esra Abaci Turk[2], P. Ellen Grant[2,3], Polina Golland[1,4], and Elfar Adalsteinsson[1,5]

[1] Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA
junshen@mit.edu
[2] Fetal-Neonatal Neuroimaging and Developmental Science Center, Boston Children's Hospital, Boston, MA, USA
[3] Harvard Medical School, Boston, MA, USA
[4] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
[5] Institute for Medical Engineering and Science, MIT, Cambridge, MA, USA

Abstract. Fetal brain MRI is useful for diagnosing brain abnormalities but is challenged by fetal motion. The current protocol for T2-weighted fetal brain MRI is not robust to motion so image volumes are degraded by inter- and intra- slice motion artifacts. Besides, manual annotation for fetal MR image quality assessment are usually time-consuming. Therefore, in this work, a semi-supervised deep learning method that detects slices with artifacts during the brain volume scan is proposed. Our method is based on the mean teacher model, where we not only enforce consistency between student and teacher models on the whole image, but also adopt an ROI consistency loss to guide the network to focus on the brain region. The proposed method is evaluated on a fetal brain MR dataset with 11,223 labeled images and more than 200,000 unlabeled images. Results show that compared with supervised learning, the proposed method can improve model accuracy by about 6% and outperform other state-of-the-art semi-supervised learning methods. The proposed method is also implemented and evaluated on an MR scanner, which demonstrates the feasibility of online image quality assessment and image reacquisition during fetal MR scans.

Keywords: Image quality assessment · Fetal magnetic resonance imaging (MRI) · Semi-supervised learning · Convolutional neural network (CNN)

1 Introduction

Fetal brain Magnetic Resonance Imaging (MRI) is an important tool complementing Ultrasound in diagnosing fetal brain abnormalities [2,9]. While MRI

© Springer Nature Switzerland AG 2020
A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12266, pp. 386–395, 2020.
https://doi.org/10.1007/978-3-030-59725-2_37

provides higher quality tissue contrast compared to Ultrasound [7,9], it is more vulnerable to motion artifacts because data acquisition is slow relative to the motion dynamics in the body [20]. This makes it challenging to adapt MRI for fetal imaging since fetal motion is more random and larger compared to adults [9]. The current protocol for T2-weighted fetal brain MRI attempts to mitigate motion artifacts by using time-efficient (∼500 ms) readouts per slice, such as the single-shot T2 weighted (SST2W) imaging acquisition. Due to safety constraints on the amount of allowable exposure to radio-frequency energy, there is a 1–2 delay between the acquisition of two consecutive slices in the stack, so obtaining an entire stack (∼30 slices) takes approximately 1 min. Orthogonal stacks are acquired and used to reconstruct the fetal brain volume. However, inter-slice and even intra-slice motion artifacts occur, contaminating the volume reconstruction [17]. Therefore, in order to improve the quality, entire stacks are usually reacquired several times [6,9], which is time-consuming. Prospective detection and reacquisition of low quality slices are expected to improve both the reconstruction quality of the brain volume, as well as the efficiency of MR scans.

Several prior studies have demonstrated the potential of Convolutional Neural Networks (CNNs) for fast image quality assessment (IQA) of MRI. Esses et al. [1] trained a CNN for volume quality assessment of T2-weighted liver MRI. Sujit et al. [15] proposed an ensemble learning method for volume quality assessment of pediatric and adult brain MRI by using multiple CNNs. However, several differences exist between these problems and fetal brain MR IQA. Specifically, these works aim to evaluate the quality of the entire stack of images instead of a single slice. Furthermore, in fetal MRI, motion is a dominant source of artifacts, typically appearing as blurs and nonuniform signal voids. Although motion is also a major source of artifacts in liver, adult brain, and cardiac MRI [12], their manifestations are different, as in this applications the motion is more regular and smaller in range compared to the motion observed in the fetus [9].

Since labeling large-scale medical image datasets is usually difficult and time-consuming, numerous semi-supervised learning methods have been proposed to leverage information in unlabeled data to improve the performance and robustness of deep neural networks. One general technique of semi-supervised learning is to infer pseudo labels from partially labeled data, such as self-training [19] and label propagation methods [4]. To yield better pseudo labels, recent methods use an ensemble of multiple neural networks which is known as self-ensembling, including temporal ensembling [8] and mean teacher [16]. In temporal ensembling, for each sample, the exponential moving average of classification outputs at different training epochs are computed and used as pseudo labels. The mean squared error (MSE) between model predictions and the pseudo labels is used as a consistency loss. One drawback of the temporal ensembling method is that it needs to keep track of the pseudo labels which is memory-consuming for large datasets. To address this problem, mean teacher method is proposed, which instead of using an ensemble of network outputs, aggregates the parameters of networks at different training step to build a teacher model. The system consists of two models with the same architecture, i.e., student and teacher. The student

model is updated with gradient during training, while the teacher model is the exponential moving average of the student model. The prediction of the teacher model is considered as pseudo label, and a consistency loss similar to temporal ensembling, is enforced between the predictions of student and teacher models. The consistency loss in self-ensembling method can also be interpreted as a regularization that smooths the network around unlabeled data. Following this interpretation, Miyato et al. proposed virtual adversarial training (VAT) [10] where they enforced consistency between predictions of original images and corresponding adversarial samples. These semi-supervised methods have also found their way into application of medical imaging, such as nuclei classification [14] and gastric diseases diagnosis [13].

In this work, we proposed a novel semi-supervised learning method for fetal MRI quality assessment. Our method extends the mean teacher model by introducing a region-of-interest (ROI) consistency for fetal brain, which let the network focus on the fetal brain ROI during feature extraction, and thus improves the accuracy of detecting non-diagnostic MR images. Evaluation showed that our method outperformed other state-of-the-art semi-supervised methods. We also implemented and evaluated the proposed method on a MR scanner, demonstrating the feasibility of online image quality assessment and image reacquisition during fetal MR scans.

2 Methods

2.1 Mean Teacher Model

In semi-supervised learning, let $\{x_1, x_2, ..., x_{N_l}\}$ be the labeled dataset with labels $\{y_1, y_2, ..., x_{N_l}\}$ and let $\{x_{N_l+1}, x_{N_l+2}, ..., x_N\}$ be the unlabeled dataset. The mean teacher model [16] consists of two networks with the same architecture, i.e., student network and teacher network, whose parameters are denoted as θ and θ' respectively.

During training, the student network is updated by minimizing the following loss function:

$$\begin{aligned} L_{\mathrm{MT}} &= L_{\mathrm{cls}} + \lambda L_{\mathrm{con}} \\ &= \frac{1}{N}\sum_{i=1}^{N_l} H(y_i, f_\theta(x_i, \eta)) + \frac{\lambda}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}(f_{\theta'}(x_i, \eta') || f_\theta(x_i, \eta)) \end{aligned} \tag{1}$$

The first term is the classification loss for labeled data, which is the cross entropy between student network prediction $f_\theta(x_i, \eta)$ and label y_i. The second term is the consistency loss between predictions of student and teacher networks. Inspired by VAT [10], we use Kullback–Leibler (KL) divergence to measure the distance between the student and teacher predictions, instead of MSE as used in the original mean teacher method [16], where η and η' denote the noise perturbation for the two networks and λ is the weight of consistency loss. The teacher network is updated as follows: $\theta'_{t+1} = \alpha\theta'_t + (1-\alpha)\theta_t$, where α is the coefficient and t is training step.

2.2 Brain ROI Consistency

In fetal brain MRI, the brain occupies a small portion of the image due to imaging parameter constraints [2]. However, the fetal brain is the ROI relevant for fetal brain MRI IQA since only the artifacts occurring in the brain affect diagnostic quality of the image. Therefore, it is essential to train the model to focus on features within the brain ROI. To fulfill this goal, We propose an ROI consistency loss to regularize the network. The overall architecture of the proposed mean teacher model with brain ROI consistency is shown in Fig. 1.

First, we introduce an ROI extraction module (Fig. 1A). For each image x, it produces a brain ROI mask R. $x_R = x \odot R$ is the masked image, where $\odot$ is the Hadamard product. The implementation of ROI extraction relies on a segmentation model. We utilize a trained U-Net in [11] to segment fetal brains from MR slices. However, since the segmentation network is trained on images with different acquisition parameters, it may yield inaccurate segmentation masks and fail to detect the brain ROI for some slices in our dataset. To improve robustness of ROI detection, instead of using the output of segmentation network directly, we aggregate the masks of images belonging to the same scan to generate a single ROI mask for the whole stack of images. The proposed algorithm is described in Fig. 1C. A stack of images are fed into the pretrained network to generate raw masks. For each mask M_i in the stack, its area A_i, center q_i and radius r_i are computed. We exclude those masks with area less than a threshold $A_{\min}$, which are assumed to be inaccurate, and let $B = \{i | 1 \leq i \leq S, A_i \geq A_{\min}\}$ be the set of remaining slices. We then compute the area-weighted mean and variance of the centers over B, i.e., $q = \frac{1}{|B|} \sum_{i \in B} A_i q_i$ and $\sigma^2 = \frac{1}{|B|} \sum_{i \in B} A_i ||q_i - q||_2^2$. The final ROI mask R is defined as the circle centered at q with radius $r = \sigma + \max_{i \in B} r_i$.

The goal of ROI consistency loss is to make the network focus on brain ROI. Let z be the output feature of the last convolution layer. $z_{\theta'}(x_i \odot R_i, \eta)$ is the feature of ROI extracted by the teacher network and $z_\theta(x_i, \eta)$ is the feature of the original image extracted by the student network. We want these features to be close to each other, so that the student can learn to detect the brain ROI from the whole image. The ROI consistency loss are defined as the MSE between these two features:

$$L_{\text{con-roi}} = \frac{1}{N} \sum_{i=1}^{N} ||z_{\theta'}(x_i \odot R_i, \eta) - z_\theta(x_i, \eta)||_2^2 \tag{2}$$

The ROI consistency loss use the feature of masked images extracted by the teacher network as reference. To guide the teacher network to learn meaningful features from the masked images, the classification loss for masked images in the labeled dataset is used as a regularization which is denoted as $L_{\text{cls-roi}}$.

$$L_{\text{cls-roi}} = \frac{1}{N} \sum_{i=1}^{N_l} H(y_i, f_\theta(x_i \odot R_i, \eta)) \tag{3}$$

We also adopted conditional entropy as an additional loss:

$$L_{\text{ent}} = H(y|x) = \frac{1}{N}\sum_{i=1}^{N} H(f_\theta(x_i, \eta), f_\theta(x_i, \eta)) \tag{4}$$

which is able to exaggerate the prediction of the network on each data point [10]. Therefore, the total loss of the proposed method is as follows.

$$L = L_{\text{cls}} + L_{\text{cls-roi}} + \lambda L_{\text{con}} + \beta L_{\text{con-roi}} + \gamma L_{\text{ent}} \tag{5}$$

where λ, β and γ are weight coefficients.

At the first couple of epochs, the teacher network cannot provide a reliable guide to the student network. For this reason, we use a ramp-up function $w(t) = \exp[-5(1-\min(t,T)/T)^2]$ for coefficients λ, β and γ, where t is the current epoch and $T = 5$.

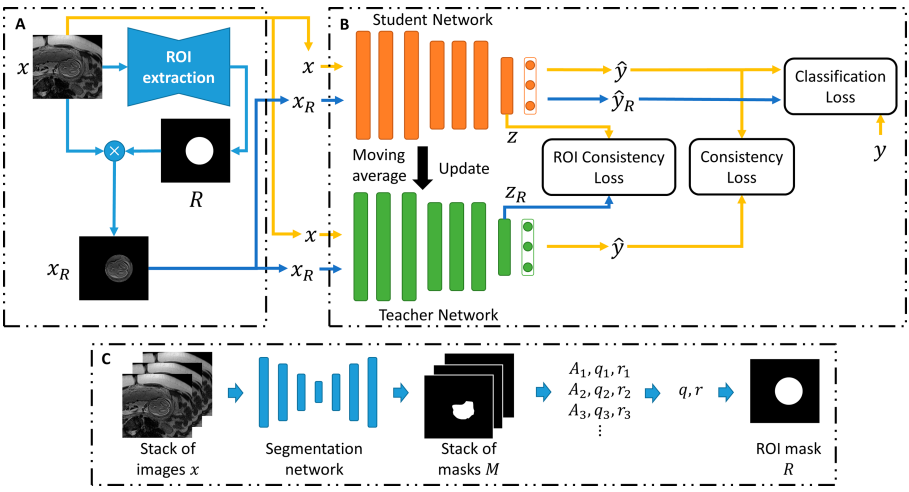

Fig. 1. Overview of the proposed method. A) Brain ROI extraction. B) Mean teacher model with ROI consistency loss. C) Details of ROI extraction algorithm.

3 Experiments and Results

3.1 Dataset

A total of 217129 images were obtained from 644 previously acquired research and clinical scans of mothers with singleton pregnancies and no pathologies, ranging in gestational age between 19 to 37 weeks. Scans were conducted at Boston Children's Hospital with Institutional Review Board approval. Scans were acquired using the SST2W sequence with median echo time TE = 115 ms, repetition time TR = 1.6 s, field of view 31 cm, and voxel size of $1.2 \times 1.2 \times 3\,\text{mm}^3$.

A set of 11223 images from 42 subjects are selected as labeled set and classified into three categories: diagnostic (D), non-diagnostic (N) and images without brain region of interest (W). Diagnostic images were characterized by sharp brain boundaries while non-diagnostic images were characterized by artifacts that occlude such features (Fig. 2). Motion artifacts manifest as signal void and blurring over the brain region. Other artifacts manifest as aliasing or the fetus not being in the field of view. A research assistant trained under radiologists labeled the dataset. The labeled dataset is divided into training (7717 images), validation (1782 images), and test (1724 images) set, where the test set consists of subjects different from training and validation sets.

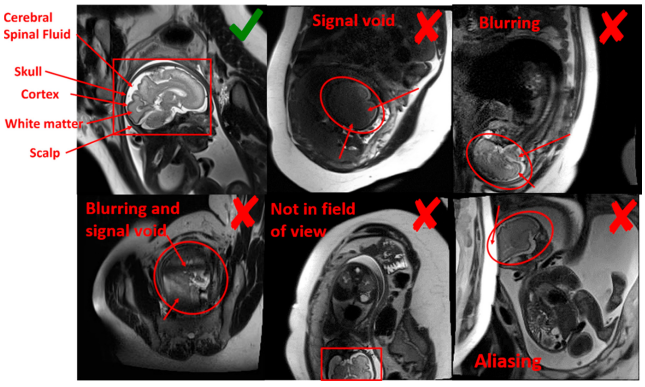

Fig. 2. Representative examples of diagnostic and nondiagnostic quality fetal brain MRI

3.2 Experiments Setup

We adopted ResNet-34 [3] as the backbone for student and teacher networks and set $\alpha = 0.994$ and $\lambda = \beta = \gamma = 1$, unless otherwise stated.

To evaluate the proposed method, we compare it with other three methods, including supervised learning, mean teacher (MT) [16] and virtual adversarial training (VAT) [10]. In addition to accuracy, we also adopted area under the ROC curves (AUC) for non-diagnostic images (N) as performance metric, since in clinical practice we are interested in detecting non-diagnostic slices and reacquiring them during MR scan.

For each method, we train the model using 1000, 2000, 4000 and all labeled data in training set (7717). For semi-supervised method, all unlabeled data are used for training. We used a batch size of 384. To balance the number of labeled and unlabeled data seen by the model, in each batch, 96 images are drawn from labeled dataset while the remains are unlabeled data. We run each experiment for 5 times and report the mean and standard deviation of evaluation metrics.

All Neural networks were implemented with PyTorch and trained on a server with an Intel Xeon E5-1650 CPU, 128 GB RAM and four NVIDIA TITAN X

GPUs. Adam [5] optimizer is used with an initial learning rate of 5×10^{-3}, and cosine learning rate decay.

3.3 Results

Results of accuracy and AUC are reported in Table 1. Results show that, the proposed method outperforms other state-of-the-art semi-supervised learning method in terms of both accuracy and AUC of non-diagnostic image. Additionally, comparing with supervised learning, the proposed approach increases accuracy and AUC by 5.82% and 0.084 respectively by learning extra information of large scale unlabeled dataset. We can also see that for smaller labeled training set (e.g., 1000 labels) the gain in accuracy from unlabeled data is higher. Besides, ablation studies were performed by setting λ, β or γ to zero to evaluate the contribution for each regularization. Results show that all the three regularization terms in our method can improve the performance of network.

Table 1. Accuracy $\pm$ std (%) and AUC for non-diagnostic image $\pm$ std over 5 runs.

Metric	Method	1000 labels	2000 labels	4000 labels	All labels
Acc.	Supervised	75.58 ± 0.93	77.40 ± 0.68	79.33 ± 0.94	79.37 ± 0.38
	VAT [10]	76.06 ± 2.18	77.51 ± 1.90	80.45 ± 2.35	81.25 ± 1.21
	MT [16]	79.27 ± 0.85	80.35 ± 0.56	81.21 ± 0.81	81.89 ± 0.63
	Proposed	$\mathbf{82.87 \pm 0.92}$	$\mathbf{83.73 \pm 0.86}$	$\mathbf{84.37 \pm 0.37}$	$\mathbf{85.19 \pm 0.19}$
	$\lambda = 0$	80.88 ± 1.07	81.38 ± 0.70	82.47 ± 0.42	82.88 ± 0.31
	$\beta = 0$	80.78 ± 0.66	82.01 ± 1.03	83.27 ± 0.34	83.81 ± 0.52
	$\gamma = 0$	80.61 ± 0.26	80.92 ± 0.61	82.68 ± 0.53	83.77 ± 0.40
AUC	Supervised	0.788 ± 0.016	0.818 ± 0.012	0.826 ± 0.008	0.815 ± 0.012
	VAT [10]	0.815 ± 0.021	0.822 ± 0.014	0.833 ± 0.017	0.844 ± 0.044
	MT [16]	0.831 ± 0.008	0.851 ± 0.005	0.856 ± 0.011	0.864 ± 0.006
	Proposed	$\mathbf{0.869 \pm 0.008}$	$\mathbf{0.881 \pm 0.003}$	$\mathbf{0.889 \pm 0.007}$	$\mathbf{0.899 \pm 0.006}$
	$\lambda = 0$	0.829 ± 0.007	0.822 ± 0.001	0.841 ± 0.011	0.854 ± 0.008
	$\beta = 0$	0.854 ± 0.006	0.872 ± 0.005	0.875 ± 0.004	0.887 ± 0.006
	$\gamma = 0$	0.855 ± 0.009	0.860 ± 0.003	0.878 ± 0.006	0.882 ± 0.005

3.4 Online Implementation

To further evaluation the proposed method and its performance in clinical practice, we developed and implemented a pipeline that runs the IQA CNN during fetal MR scans to assign a IQA score to each slice and reacquire those slices with low IQA scores. The trained CNN is deployed on a GPU (NVIDIA 1050Ti) equipped computer which is connected to the scanner's internal network. In each scan, N_{acq} slices were acquired and the IQA scores are computed as $s = 1 - P_N$, where P_N is probability of non-diagnostic image. Then the N_{re} slices with lowest IQA scores were reacquired. The proportion of re-acquisition is denoted as

$q = N_{re}/N_{acq}$. We performed a simulation study on the test set consisting of stacks of images with 20 to 40 slices where about one third of the images are of low quality in average (in the worst case, over 60% of slices in a stack are contaminated by motion artifacts). The number of missing non-diagnostic images is shown in Fig. 3a, where 'random' means random re-acquisition. The proposed method outperforms the supervised baseline and only misses one non-diagnostic slice in average when $q = 50\%$.

For in vivo study, fetal scans were performed on a 3T MR scanner with $N_{acq} = 20, q = 0.5$. Figure 3b shows 4 images from 3 separate scans, where the originally acquired slices (top row) were motion degraded, and the re-acquired ones (bottom row) were not. These results demonstrated the feasibility of online detection of non-diagnostic MR images during fetal scans using the proposed deep learning method.

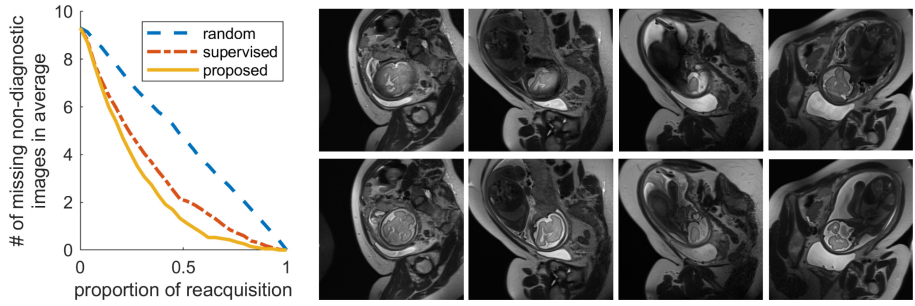

Fig. 3. (a) Number of non-diagnostic slices that are not detected by the IQA pipeline. (b) Four examples from three separate in vivo scans showing motion artifacts in the originally acquired images (top row), and much cleaner images when the same slice locations were re-acquired (bottom row).

4 Conclusions

In this paper, we proposed a novel semi-supervised learning method for fetal MRI quality assessment. Our method extend the mean teacher model by introducing a ROI consistency for fetal brain which let the network focus on brain ROI during feature extraction and therefore improve the accuracy of detecting non-diagnostic MR images. Evaluation showed that our method outperformed other state-of-the-art semi-supervised methods as well. We also implemented and evaluated the proposed method on a MR scanner, demonstrating the feasibility of online image quality assessment and image requisition during fetal MR scans, which can work in tandem with fetal motion tracking algorithm [18] to improve image quality as well as efficiency of imaging workflow.

References

1. Esses, S.J., et al.: Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture. J. Magn. Reson. Imaging **47**(3), 723–728 (2018)
2. Gholipour, A., et al.: Fetal MRI: a technical update with educational aspirations. Concepts Magn. Reson. Part A **43**(6), 237–266 (2014)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
4. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5070–5079 (2019)
5. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Kline-Fath, B., Bahado-Singh, R., Bulas, D.: Fundamental and Advanced Fetal-imaging: Ultrasound and MRI. Lippincott Williams & Wilkins, Philadelphia (2014)
7. Kul, S., et al.: Contribution of MRI to ultrasound in the diagnosis of fetal anomalies. J. Magn. Reson. Imaging **35**(4), 882–890 (2012)
8. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
9. Malamateniou, C., et al.: Motion-compensation techniques in neonatal and fetal MR imaging. Am. J. Neuroradiol. **34**(6), 1124–1136 (2013)
10. Miyato, T., Maeda, S.I., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1979–1993 (2018)
11. Salehi, S.S.M., et al.: Real-time automatic fetal brain extraction in fetal MRI by deep learning. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 720–724. IEEE (2018)
12. Schreiber-Zinaman, J., Rosenkrantz, A.B.: Frequency and reasons for extra sequences in clinical abdominal MRI examinations. Abdom. Radiol. **42**(1), 306–311 (2017)
13. Shang, H., et al.: Leveraging other datasets for medical imaging classification: evaluation of transfer, multi-task and semi-supervised learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11768, pp. 431–439. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32254-0_48
14. Su, H., Shi, X., Cai, J., Yang, L.: Local and global consistency regularized mean teacher for semi-supervised nuclei classification. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 559–567. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_62
15. Sujit, S.J., Coronado, I., Kamali, A., Narayana, P.A., Gabr, R.E.: Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. J. Magn. Reson. Imaging **50**, 1260–1267 (2019)
16. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, pp. 1195–1204 (2017)
17. Tourbier, S., Bresson, X., Hagmann, P., Thiran, J.P., Meuli, R., Cuadra, M.B.: An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization. NeuroImage **118**, 584–597 (2015)

18. Xu, J., et al.: Fetal pose estimation in volumetric MRI using a 3D convolution neural network. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 403–410. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_44
19. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189–196 (1995)
20. Zaitsev, M., Maclaren, J., Herbst, M.: Motion artifacts in MRI: a complex problem with many partial solutions. J. Magn. Reson. Imaging **42**(4), 887–901 (2015)