UM-CAM: Uncertainty-weighted Multi-resolution Class Activation Maps for Weakly-supervised Fetal Brain Segmentation

Jia Fu[1], Tao Lu[2], Shaoting Zhang[1,3], and Guotai Wang[1,3](✉)

[1] University of Electronic Science and Technology of China, Chengdu 611731, China
[2] Sichuan Provincial People's Hospital, Chengdu 610072, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai 200030, China
guotai.wang@uestc.edu.cn

Abstract. Accurate segmentation of the fetal brain from Magnetic Resonance Image (MRI) is important for prenatal assessment of fetal development. Although deep learning has shown the potential to achieve this task, it requires a large fine annotated dataset that is difficult to collect. To address this issue, weakly-supervised segmentation methods with image-level labels have gained attention, which are commonly based on class activation maps from a classification network trained with image-level labels. However, most of these methods suffer from incomplete activation regions, due to the low-resolution localization without detailed boundary cues. To this end, we propose a novel weakly-supervised method with image-level labels based on semantic features and context information exploration. We first propose an Uncertainty-weighted Multi-resolution Class Activation Map (UM-CAM) to generate high-quality pixel-level supervision. Then, we design a Geodesic distance-based Seed Expansion (GSE) method to provide context information for rectifying the ambiguous boundaries of UM-CAM. Extensive experiments on a fetal brain dataset show that our UM-CAM can provide more accurate activation regions with fewer false positive regions than existing CAM variants, and our proposed method outperforms state-of-the-art weakly-supervised segmentation methods learning from image-level labels.

Keywords: Weakly-supervised segmentation · Class activation map · Geodesic distance · Fetal MRI

1 Introduction

Brain extraction is the first step in fetal Magnetic Resonance Image (MRI) analysis in advanced applications such as brain tissue segmentation [15] and quantitative measurement [22,24], which is essential for assessing fetal brain development and investigate the neuroanatomical correlation of cognitive impairments [14]. Current research based on Convolutional Neural Network (CNN) [5,19] has achieved promising performance for automatic fetal brain extraction from pixel-wise annotated fetal MRI. However, it is labor-intensive, time-consuming, and

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 315–324, 2023.
https://doi.org/10.1007/978-3-031-43990-2_30

expensive to collect a large-scale pixel-wise annotated dataset, especially for images with poor quality and large variations. To address these issues, weakly-supervised segmentation methods with image-level supervision [21] are introduced due to their minimal annotation demand. However, learning from image-level supervision is extremely challenging since the image-level label only provides the existence of object class, but cannot indicate the information about location and shape that are essential for the segmentation task [1].

Prevailing methods learning from image-level labels for segmentation commonly produce a coarse localization of the objects based on Class Activation Maps (CAM) [27]. Due to the weak annotation, the CAMs from the classification network can only provide rough localization and coarse boundaries of objects. To alleviate the problem, a lot of approaches have been proposed, which can be categorized as one-stage and two-stage methods. One-stage methods aim to generate pixel-level segmentation by training a segmentation branch simultaneously with a classification network. For example, Reliable Region Mining (RRM) [26] comprises two parallel branches, in which pixel-level pseudo masks are produced from the classification branch and refined by Conditional Random Field (CRF) to supervise the segmentation branch. Despite their efficiency, one-stage methods commonly achieve inferior segmentation accuracy and incomplete activation of targets, owing to the failure to capture detailed contextual information from image-level labels [2,7].

In contrast to one-stage methods, two-stage methods can perform favourably, as they leverage dense labels generated by the classification network to train a segmentation network [12]. For instance, Discriminative Region Suppression (DRS) [9] suppresses the attention on discriminative regions and expands it to adjacent less activated regions. However, these methods leverage the CAMs from the deep layer of the classification network and raise the inherent drawback, i.e., low resolution, leading to limited localization and smooth boundaries of objects. Han et al. [8] proposed multi-layer pseudo supervision to reduce the false positive rate in segmentation results, while the weights for pseudo masks from different layers are constants that cannot be adaptive. Besides, though the quality of CAMs improves, they are still insufficient to provide accurate object boundaries for segmentation. Numerous methods [6,10,11] have been proposed to explore boundary information. For example, Kolesnikov et al. [10] proposed a joint loss function that constrains the global weighted rank pooling and low-level object boundary to expand activation regions. AffinityNet [1] trains another network to learn the semantic similarity between pixels and then propagates the semantics to adjacent pixels via random walk. Nevertheless, these methods use the initial seeds generated from the CAM method, resulting in limited performance when the object-related seeds from CAM are small and sparse. Thus, improving the initial prediction and exploring boundary information are both important for accurate object segmentation.

In this work, we propose a novel weakly-supervised method for accurate fetal brain segmentation using image-level labels. Our contribution can be summarized as follows: 1) We design an Uncertainty-weighted Multi-resolution CAM

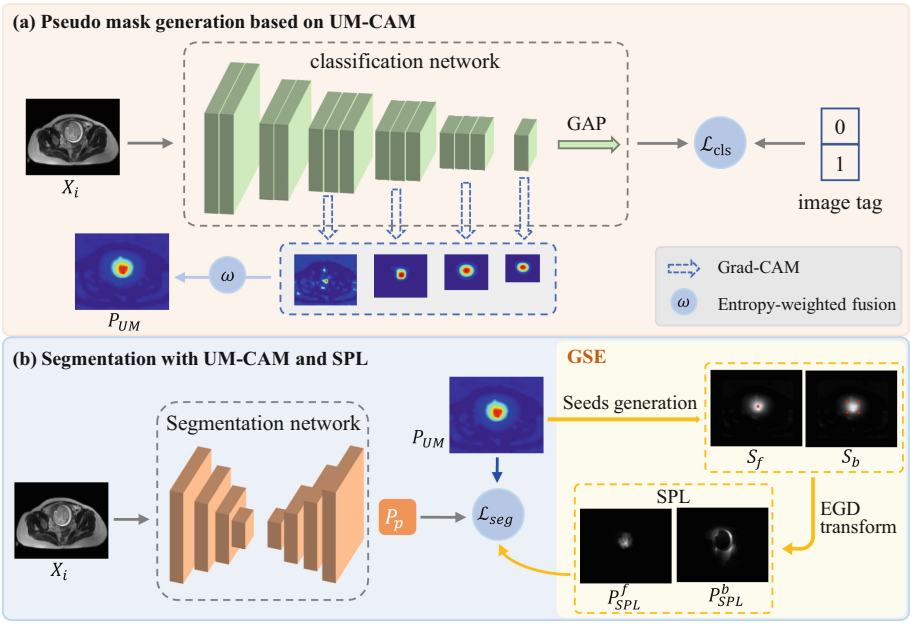

Fig. 1. Overview of the proposed method. (a) Uncertainty-weighted Multi-resolution CAM (UM-CAM) obtained by a classification network, (b) Segmentation model trained with UM-CAM and auxiliary supervision from Seed-derived Pseudo Label (SPL).

(UM-CAM) to integrate low- and high-resolution CAMs via entropy weighting, which can leverage the semantic features extracted from the classification network adaptively and eliminate the noise effectively. 2) We propose a Geodesic distance-based Seed Expansion (GSE) method to generate Seed-derived Pseudo Labels (SPLs) that can provide boundary cues for training a better segmentation model. 3) Extensive experiments conducted on a fetal brain dataset demonstrate the effectiveness of the proposed method, which outperforms several state-of-the-art approaches for learning from image-level labels.

2 Method

An overview of our method is presented in Fig. 1. First, to obtain high-quality pseudo masks, Uncertainty-weighted Multi-resolution CAM (UM-CAM) is produced by fusing low- and high-resolution CAMs from different layers of the classification network. Second, seed points are obtained from the UM-CAM automatically, and used to generate Seed-derived Pseudo Labels (SPL) via Geodesic distance-based Seed Expansion (GSE). The SPL provides more detailed context information in addition to UM-CAM for training the final segmentation model.

2.1 Psuedo Mask Generation Based on UM-CAM

Initial Response via Grad-CAM. A typical classification network consists of convolutional layers as a feature extractor, followed by global average pooling and a fully connected layer as the output classifier [3]. Given a set of training images, the classification network is trained with class labels. After training, the Grad-CAM method is utilized to compute the weights α_k for the k-th channel of a feature map f at a certain layer via gradient backpropagation from the output node for the foreground class. The foreground activation map A can be obtained from a weighted combination of feature maps and followed by a ReLU activation [20], which is formulated as:

$$\alpha_k = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial y}{\partial f_k(i)}, \tag{1}$$

$$A(i) = ReLU(\sum_k \alpha_k f_k(i)), \tag{2}$$

where y is classification prediction score for the foreground. i is pixel index, and N is the pixel number in the image.

Multi-resolution Exploration and Integration. The localization map for each image typically provides discriminative object parts, which is insufficient to provide supervision for the segmentation task. As shown in Fig. 1(a), the activation maps generated from the shallow layers of the classification network contain high-resolution semantic features but suffer from noisy and dispersive localization. In contrast, activation maps generated from the deeper layers perform smoother localization but lack high-resolution information. To take advantage of activation maps from shallow and deep layers, we proposed UM-CAM to integrate the multi-resolution CAMs by uncertainty weighting. Let us denote a set of activation maps from M convolutional blocks as $\mathcal{A} = \{A_m\}_{m=0}^{M}$. Each activation map is interpolated to the input size and normalized by its maximum to the range of [0,1], and the normalized activation maps are $\hat{\mathcal{A}} = \{\hat{A}_m\}_{m=0}^{M}$. To minimize the uncertainty of pseudo mask, UM-CAM integrates the confident region of multi-resolution CAMs adaptively, which can be presented as the entropy-weighted combination of CAMs:

$$w_m(i) = 1 - (- \sum_{j=(b,f)} \hat{A}_m^j(i) log \hat{A}_m^j(i)), \tag{3}$$

$$P_{UM}(i) = \frac{\sum_m w_m(i) \hat{A}_m(i)}{\sum_m w_m(i)}, \tag{4}$$

where $\hat{A}_m^b$ and $\hat{A}_m^f$ represent the background and foreground probability of $\hat{A}_m$, respectively. w_m is the weight map for $\hat{A}_m$, and P_{UM} is the UM-CAM for the target.

2.2 Robust Segmentation with UM-CAM and SPL

Though UM-CAM is better than the CAM from the deep layer of the classification network, it is still insufficient to provide accurate object boundaries that are important for segmentation. Motivated by [13], we propose a Geodesic distance-based Seed Expansion (GSE) method to generate Seed-derived Pseudo Label (SPL) that contains more detailed context information. The SPL is combined with UM-CAM to supervise the segmentation model, as shown in Fig. 1 (b).

Concretely, we adopt the centroid and the corner points of the bounding box obtained from UM-CAM as the foreground seeds S_f and background seeds S_b, respectively. To efficiently leverage these seed points, SPL is generated via Exponential Geodesic Distance (EGD) transform of the seeds, leading to a foreground cue map P^f_{SPL} and a background cue map P^b_{SPL}. The values of P^b_{SPL} and P^f_{SPL} represent the similarity between each pixel and background/foreground seed points, which can be computed as:

$$P^b_{SPL}(i) = e^{-\alpha \cdot D_b(i)}, \quad P^f_{SPL}(i) = e^{-\alpha \cdot D_f(i)}, \tag{5}$$

$$D_b(i) = \min_{j \in S_b} D_{geo}(i, j, I), \quad D_f(i) = \min_{j \in S_f} D_{geo}(i, j, I), \tag{6}$$

$$D_{geo}(i, j, I) = \min_{p \in P_{i,j}} \int_0^1 \| \bigtriangledown I(p(n)) \cdot u(n) \| \, dn \tag{7}$$

where $P_{i,j}$ is the set of all paths between pixels i and j. $D_b(i)$ and $D_f(i)$ represent the minimal geodesic distance between target pixel i and background/foreground seed points, respectively. p is one feasible path and it is parameterized by $n \in [0, 1]$. $u(n) = p'(n) / \|p'(n)\|$ is a unit vector that is tangent to the direction of the path.

Based on the supervision from UM-CAM and SPL, the segmentation network can be trained by minimizing the following joint object function:

$$L_{seg} = \lambda L_{CE}(P_p, P_{UM}) + (1 - \lambda) L_{CE}(P_p, P_{SPL}). \tag{8}$$

where P_p is the prediction of the segmentation network, and λ is a weight factor to balance the supervision of UM-CAM and SPL. L_{CE} is the Cross-Entropy (CE) loss.

3 Experiments and Results

3.1 Experimental Details

Dataset. We collected clinical T2-weighted MRI data of 115 pregnant women in the second trimester with Single-shot Fast Spin-echo (SSFSE). The data were acquired in axial view with pixel size between 0.5547 mm × 0.5547 mm and 0.6719 mm × 0.6719 mm and slice thickness between 6.50 mm and 7.15 mm. Each slice was resampled to a uniform pixel size of 1 mm × 1 mm. In all experiments,

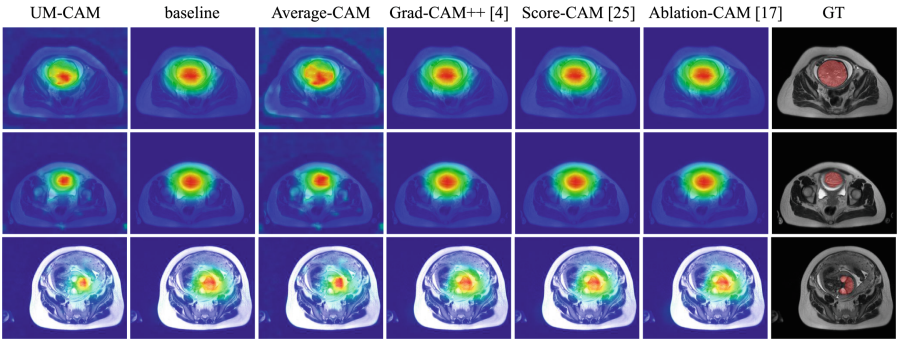

Fig. 2. Visual comparison between CAMs obtained by different methods.

we used 80 volumes with 976 positive and 3140 negative slices for training, 10 volumes with 116 positive and 408 negative slices for validation, and 25 volumes with 318 positive and 950 negative slices for testing. Positive and negative slice mean containing the brain and not, respectively. The ground truth was manually annotated by radiologists. Note that we used image-level labels for training and pixel-level ground truth for validation and testing.

Implementation Details. To boost the generalizability, we applied spatial and intensity-based data augmentation during the training stage, including gamma correction, random rotation, random flipping, and random cropping. For 2D classification, we employed VGG-16 [23] as backbone architecture, in which an additional convolutional layer is used to substitute the last three fully connected layers. The classification network was trained with 200 epochs using CE loss. Stochastic Gradient Descent (SGD) optimizer was used for training with batch size 32, momentum 0.99, and weight decay 5×10^{-4}. The learning rate was initialized to 1×10^{-3}. We used UNet [18] as the segmentation network. The learning rate was set as 0.01, and the SGD optimizer was used for training with 200 epochs, batch size 12, momentum 0.9, and weight decay 5×10^{-4}. The hyper-parameter M and λ were set as 4 and 0.1 based on the best results on the validation set, respectively. We used Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD_{95}) to evaluate the quality of 2D pseudo masks and the final segmentation results in 3D space.

3.2 Ablation Studies

Stage1: Quality of Pseudo Masks Obtained by UM-CAM. To evaluate the effectiveness of UM-CAM, we compared different pseudo mask generation strategies: 1) Grad-CAM (baseline): only using CAMs from the last layer of the classification network generated by using Grad-CAM method [20], 2) Average-CAM: fusing multi-resolution CAMs via averaging, 3) UM-CAM: fusing multi-resolution CAMs via uncertainty weighting. Table 1 lists the quantitative

Table 1. Ablation study on the validation set to validate the effectiveness of UM-CAM and SPL. * denotes p-value < 0.05 when comparing with the second place method.

Method		DSC (%)	HD_{95} (pixels)
Pseudo mask generation	Grad-CAM (baseline)	74.48 ± 13.26	39.88 ± 14.66
	Average-CAM	77.40 ± 13.28	38.70 ± 19.75
	UM-CAM	$\mathbf{79.00 \pm 12.83}$*	$\mathbf{35.33 \pm 16.98}$
Segmentation	Grad-CAM (baseline)	78.69 ± 10.02	22.17 ± 16.60
	UM-CAM	85.22 ± 6.62	5.26 ± 4.23
	SPL	89.05 ± 4.30	3.84 ± 4.07
	UM-CAM+SPL	$\mathbf{89.76 \pm 5.09}$*	$\mathbf{3.10 \pm 2.61}$

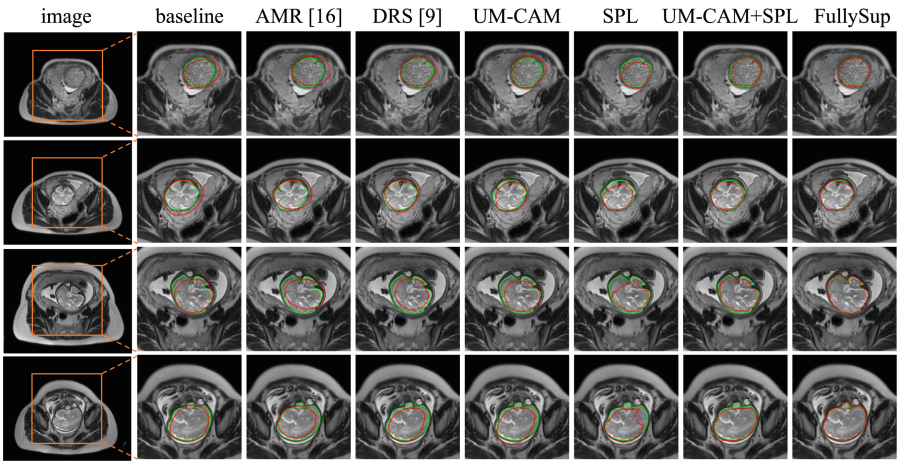

Fig. 3. Visual comparison of our method and other weakly-supervised segmentation methods. The green and red contours indicate the boundaries of ground truths and segmentation results, respectively. (Color figure online)

evaluation results of these methods, in which the segmentation is converted from CAMs using the optimal threshold found by grid search method. It can be seen that when fusing the information from multiple convolutional layers, the quality of pseudo masks improves. The proposed UM-CAM improves the average DSC by 4.52% and 1.60% compared with the baseline and Average-CAM, respectively. Figure 2 shows a visual comparison between CAMs obtained by the different methods. It can be observed that there are fewer false positive activation regions of UM-CAM compared with the other methods.

Stage2: Training Segmentation Model with UM-CAM and SPL. To investigate the effectiveness of SPL, we compared it with several segmentation models: 1) Grad-CAM (baseline): only using the pseudo mask generated from

Table 2. Comparison between ours and existing weakly-supervised segmentation methods. * denotes p-value < 0.05 when comparing with the second place weakly supervised method.

Method	Validation set		Test set	
	DSC (%)	HD_{95} (pixels)	DSC (%)	HD_{95} (pixels)
Grad-CAM++ [4]	74.52 ± 13.29	39.87 ± 14.69	76.60 ± 10.58	37.49 ± 11.72
Score-CAM [25]	74.49 ± 13.35	39.88 ± 14.83	76.58 ± 10.60	37.54 ± 11.73
Ablation-CAM [17]	74.56 ± 13.29	39.76 ± 14.64	76.55 ± 10.61	37.57 ± 11.82
AMR [16]	78.77 ± 8.83	11.53 ± 9.82	79.79 ± 6.86	11.35 ± 8.02
DRS [9]	84.98 ± 5.62	7.17 ± 8.01	83.79 ± 7.81	7.06 ± 5.13
UM-CAM+SPL (ours)	$\mathbf{89.76 \pm 5.09}$*	$\mathbf{3.10 \pm 2.61}$*	$\mathbf{90.22 \pm 3.75}$*	$\mathbf{4.04 \pm 4.26}$*
FullySup	95.98 ± 3.17	1.22 ± 0.65	96.51 ± 2.67	1.10 ± 0.40

Grad-CAM to train the segmentation model, 2) UM-CAM: only using UM-CAM as supervision for the segmentation model, 3) SPL: only using SPL as supervision, 4) UM-CAM+SPL: our proposed method using UM-CAM and SPL supervision for the segmentation model. Quantitative evaluation results in the second section of Table 1 show that the network trained with UM-CAM and SPL supervision achieves an average DSC score of 89.76%, improving the DSC by 11.07% compared to the baseline model. Figure 3 depicts a visual comparison between these models. It shows that SPL supervision with context information can better discriminate the fetal brain from the background, leading to more accurate boundaries for segmentation.

3.3 Comparison with State-of-the-art Methods

We compared three CAM invariants with our UM-CAM in pseudo mask generation stage, including Grad-CAM++ [4], Score-CAM [25], and Ablation-CAM [17]. Table 2 shows the quantitative results of these CAM variants. It can be seen that GradCAM++, Score-CAM, and Ablation-CAM achieve similar performance, which is consistent with the visualization results shown in Fig. 2. The proposed UM-CAM achieves higher accuracy than existing CAM variants, which generates more accurate boundaries that are closer to the ground truth.

We further compared the proposed method with fully supervised method (FullySup) and two state-of-the-art weakly-supervised methods, including DRS [9] that spreads the attention to adjacent non-discriminative regions by suppressing the attention on discriminative regions, and AMR [16] that incorporates a spotlight branch and a compensation branch to dig out more complete object regions. Table 2 lists the segmentation results of these methods. Our proposed method achieves an average DSC of 90.22% and an average HD_{95} of 4.04 pixels, which is at least 3.02 pixels lower than the other weakly-supervised methods on the testing set. It indicates that the proposed method can generate segmentation results with more accurate boundaries. Figure 3 shows some

qualitative visualization results. The DRS and AMR predictions appear to be coarse and inaccurate in the boundary regions, while our proposed method generates more accurate segmentation results, even similar to those generated from the fully supervised model for some easy samples.

4 Conclusion

In this paper, we presented an uncertainty and context-based method for fetal brain segmentation using image-level supervision. An Uncertainty-weighted Multi-resolution CAM (UM-CAM) was proposed to integrate multi-resolution activation maps via uncertainty weighting to generate high-quality pixel-wise supervision. We proposed a Geodesic distance-based Seed Expansion (GSE) method to produce Seed-derived Pseudo Labels (SPL) containing detailed context information. The SPL is combined with UM-CAM for training the segmentation network. The proposed method was evaluated on the fetal brain segmentation task, and experimental results demonstrated the effectiveness of the proposed method and suggested the potential of our proposed method for obtaining accurate fetal brain segmentation with low annotation cost. In the future, it is of interest to validate our method with other segmentation tasks and apply it to other backbone networks.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (62271115).

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: CVPR, pp. 4981–4990 (2018)
2. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: CVPR, pp. 4253–4262 (2020)
3. Chang, Y.T., Wang, Q., Hung, W.C., Piramuthu, R., Tsai, Y.H., Yang, M.H.: Weakly-supervised semantic segmentation via sub-category exploration. In: CVPR, pp. 8991–9000 (2020)
4. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: WACV, pp. 839–847 (2018)
5. Ebner, M.: An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI. Neuroimage **206**, 116324 (2020)
6. Fan, J., Zhang, Z., Song, C., Tan, T.: Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: CVPR, pp. 4283–4292 (2020)
7. Gao, W., et al.: TS-CAM: Token semantic coupled attention map for weakly supervised object localization. In: ICCV, pp. 2886–2895 (2021)
8. Han, C., et al.: Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Med. Image Anal. **80**, 102487 (2022)

9. Kim, B., Han, S., Kim, J.: Discriminative region suppression for weakly-supervised semantic segmentation. In: AAAI. vol. 35, pp. 1754–1761 (2021)
10. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: ECCV, pp. 695–711 (2016)
11. Lee, S., Lee, M., Lee, J., Shim, H.: Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: CVPR, pp. 5495–5505 (2021)
12. Li, Y., Kuang, Z., Liu, L., Chen, Y., Zhang, W.: Pseudo-mask matters in weakly-supervised semantic segmentation. In: ICCV, pp. 6964–6973 (2021)
13. Luo, X., et al.: MIDeepSeg: minimally interactive segmentation of unseen objects from medical images using deep learning. Med. Image Anal. **72**, 102102 (2021)
14. Makropoulos, A., Counsell, S.J., Rueckert, D.: A review on automatic fetal and neonatal brain MRI segmentation. Neuroimage **170**, 231–248 (2018)
15. Makropoulos, A., et al.: Automatic whole brain MRI segmentation of the developing neonatal brain. IEEE Trans. Med. Imaging **33**(9), 1818–1831 (2014)
16. Qin, J., Wu, J., Xiao, X., Li, L., Wang, X.: Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In: AAAI, vol. 36, pp. 2117–2125 (2022)
17. Ramaswamy, H.G., et al.: Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization. In: ICCV, pp. 983–991 (2020)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241 (2015)
19. Salehi, S.S.M., et al.: Real-time automatic fetal brain extraction in fetal MRI by deep learning. In: ISBI, pp. 720–724 (2018)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)
21. Shen, W., et al.: A survey on label-efficient deep image segmentation: bridging the gap between weak supervision and dense prediction. IEEE Trans. Pattern Anal. Mach. Intell. **45**(8), 9284–9305 (2023)
22. Shi, W., et al.: Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty. Neuroimage **223**, 117316 (2020)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR, pp. 1–14 (2015)
24. Sridar, P., et al.: Automatic measurement of thalamic diameter in 2-D fetal ultrasound brain images using shape prior constrained regularized level sets. IEEE J. Biomed. Health Inform. **21**(4), 1069–1078 (2016)
25. Wang, H., et al.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: CVPR workshops, pp. 24–25 (2020)
26. Zhang, B., Xiao, J., Wei, Y., Sun, M., Huang, K.: Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: AAAI, vol. 34, pp. 12765–12772 (2020)
27. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR, pp. 2921–2929 (2016)