Propagating Uncertainty Across Cascaded Medical Imaging Tasks for Improved Deep Learning Inference

Raghav Mehta[1(✉)], Thomas Christinck[1], Tanya Nair[1], Paul Lemaitre[1], Douglas Arnold[2,3], and Tal Arbel[1]

[1] Centre for Intelligent Machines, McGill University, Montreal, Canada
raghav@cim.mcgill.ca
[2] Montreal Neurological Institute, McGill University, Montreal, Canada
[3] NeuroRx Research, Montreal, Canada

Abstract. Although deep networks have been shown to perform very well on a variety of tasks, inference in the presence of pathology in medical images presents challenges to traditional networks. Given that medical image analysis typically requires a sequence of inference tasks to be performed (e.g. registration, segmentation), this results in an accumulation of errors over the sequence of deterministic outputs. In this paper, we explore the premise that, by embedding uncertainty estimates across cascaded inference tasks, the final prediction results should improve over simply cascading the deterministic classification results or performing inference in a single stage. Specifically, we develop a deep learning framework that propagates voxel-based uncertainty measures (e.g. Monte Carlo (MC) dropout sample variance) across inference tasks in order to improve the detection and segmentation of focal pathologies (e.g. lesions, tumours) in brain MR images. We apply the framework to two different contexts. First, we demonstrate that propagating multiple sclerosis T2 lesion segmentation results along with their associated uncertainty measures improves subsequent T2 lesion detection accuracy when evaluated on a proprietary large-scale, multi-site, clinical trial dataset. Second, we show how by propagating uncertainties associated with a regressed 3D MRI volume as an additional input to a follow-on brain tumour segmentation task, one can improve segmentation results on the publicly available BraTS-2018 dataset.

1 Introduction

Deep learning methods have been shown to outperform other methods on a variety of medical imaging inference tasks [1–6]. However, challenges still remain in applying traditional networks to clinical tasks in the presence of focal pathologies. Given that a typical medical image analysis pipeline generally requires a sequence of inference tasks to be performed (e.g. registration, segmentation),

R. Mehta and T. Christinck—Equal contribution.

H. Greenspan et al. (Eds.): CLIP 2019/UNSURE 2019, LNCS 11840, pp. 1–10, 2019.
https://doi.org/10.1007/978-3-030-32689-0_3

errors made over the sequence of deterministic models can accumulate and affect the downstream clinical task of interest (e.g. survival prediction). For example, the reported underperformance of the popular U-Net in the detection and segmentation of very small lesions [5] is problematic in the context of Multiple Sclerosis (MS), in that detecting all lesions, including small ones, in patient MRI is important for disease staging, prognosis, and monitoring treatment efficacy. Recently, it has been shown that popular deep networks adapted for the synthesis of missing MRI sequences (e.g. FLAIR) underperform in the presence of tumours. This negatively affects the downstream tasks of tumour classification, and staging and sub-type segmentation [2,11] that rely on the presence of these sequences. In this paper, we hypothesize that the performance of the downstream tasks in a medical image analysis pipeline should improve if, in addition to deterministic predictions, uncertainty estimates are propagated across cascaded inference tasks.

Recently, Bayesian machine learning approaches have begun to address the limitation of deterministic deep learning methods by providing an uncertainty associated with each prediction. Gal and Ghahramani [7] showed that by training a neural network with dropout and taking Monte Carlo (MC) samples of the prediction using dropout at test time, one can estimate the uncertainty associated with the output of deep learning methods. MC dropout based uncertainty estimation has been used in a variety of medical imaging tasks recently [8–10], ranging from modality synthesis [2] to nodule detection [8] and lesion detection and segmentation [5]. Many of these papers report that prediction uncertainty can be used to estimate regions of an image where the network is prone to error [2,9]. Others demonstrate an improved performance when the network output is evaluated on its most certain predictions [5,10]. While these approaches illustrate how estimating uncertainty in medical imaging tasks is useful in a clinical scenario, they do not show how uncertainty can be used to inform or improve network performance on a downstream task. In [8], the authors begin to address this limitation by showing how uncertainty from a 2D lung nodule segmentation network can be used to reduce the *false positives* in a subsequent 3D detection network centered on regions of interest (ROIs). Although appropriate in the context of lung nodule detection, this is not the general case in medical imaging applications where false negative reduction is also, sometimes more so, of interest.

To this end, we develop a general deep learning framework that embeds uncertainty estimates across cascaded inference tasks in order to improve the performance of the downstream task of interest. Specifically, two different medical imaging contexts are investigated in which voxel-based uncertainty measures based on MC dropout (e.g. sample variance) are propagated to downstream tasks, where they are shown to improve network performance for the detection and segmentation of focal pathologies (e.g. lesions, tumours) in brain images by reducing both false positives *and* false negative predictions.

In the first context, a 3D fully-convolutional segmentation network is trained on a large multi-site, multi-scanner, proprietary dataset of MS patient MRI.

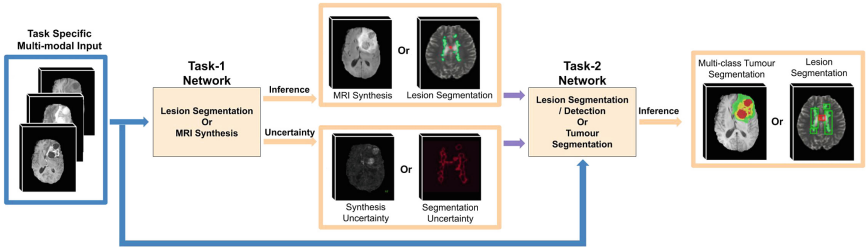


Fig. 1. Overview of the proposed general framework for propagating inference results and their associated uncertainties along sequential tasks in medical image analysis (Ex. MS T2 lesion segmentation, and MR synthesis - brain tumour segmentation). (Color figure online)

Segmentation uncertainty estimates are based on MC dropout sample variance. A second segmentation network is trained with this uncertainty as an additional input. Experimental results indicate that uncertainty propagation improves the T2 lesion true positive rate (TPR) from 0.78 to 0.84, in comparison to baseline (one stage network), at the clinically relevant false discovery rate (FDR) of 0.2. In the second context, a two-stage MR sequence synthesis and tumour segmentation pipeline is developed, which is trained and tested on the publicly available MICCAI 2018 BraTS dataset [13]. Experimental results indicate that propagating the synthesized image along with its associated uncertainty map to the downstream tumour segmentation network improves the Dice performance by anywhere from 2–10%, in comparison to only propagating synthesized image. Together, these contexts demonstrate that network uncertainty captures information supplementary to typical deterministic outputs, and can be successfully leveraged to improve the performance of downstream networks.

2 Methodology: Propagating Uncertainty Across Inference Tasks

An overview of the proposed method for propagating uncertainties across inference tasks is given in Fig. 1. The framework includes two different task specific networks, one for each of the sequential inference tasks, where each network takes images, here multi-modal MRI volumes, as inputs (blue arrows in Fig. 1). Task-1 produces the results of inference and their associated uncertainty values. In addition to MRI inputs, these outputs are provided to the Task-2 network as inputs (dark purple arrows in Fig. 1). We hypothesize that including the Task-1 network uncertainty output as an additional input to the Task-2 network will improve its performance.

The framework is general in that a number of methods can be used to estimate the uncertainties. In this work, the uncertainties produced by the Task-1 network are estimated using MC dropout sampling [7], where a network is trained using dropout and, during testing, the input is passed through the network with

dropout N times to obtain N Monte-Carlo (MC) samples. The mean of these samples is taken as the network's output prediction, and the variance of the samples is used to estimate its uncertainty. Note that both the Task-1 network and the Task-2 network are trained separately. An end-to-end training of networks is not considered here as model uncertainties (resulting from MC-Dropout) on training cases would not properly reflect the model uncertainties on unseen test cases.

This work investigates two different pipelines, each of which cascades two different inference tasks. For the first pipeline, the Task-1 network consists of a Bayesian U-Net (BU-Net) [5], a segmentation network that takes multi-modal brain MRI of MS patients and produces a T2 lesion segmentation and a voxel-level uncertainty map. For the second pipeline, the Task-1 network is a synthesis network, which takes multi-modal MR sequences of patients with tumours (e.g. T1, T1ce, T2) as input and generates an additional unavailable MR sequence (e.g. FLAIR) along with the uncertainties for the regressed volume. For this task, the multi-task Regression-Segmentation Network (RS-Net) proposed in [2] is used.

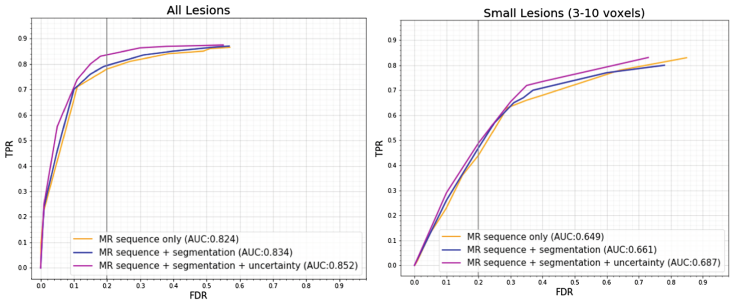
The Task-2 network for both MS and brain tumour pipelines is a segmentation network which either results in MS T2 lesion segmentation or multi-class brain tumour segmentation, respectively. A modified 3D U-Net [12] is used for this task. Like the original 3D U-Net [12], the network consists of encoder and decoder paths that contain convolution, pooling, and up-sampling/deconvolution operations. High-resolution features from the encoder path are combined with the up-sampled output of decoder in an attempt to preserve high-resolution features. Each convolution is followed by non-linear activation (Leaky ReLU/ReLU). Instead of using the batch-normalization layer used in the original U-Net, we use group [14]/instance normalization [15]. This typically improves performance for small batch sizes. For MS T2 lesion segmentation, the network is trained using a combined, equally weighted Sorensen-Dice loss and binary cross-entropy loss, and produces binary lesion segmentation output. For tumour segmentation, the network is trained using categorical cross-entropy loss and produces multi-class tumour segmentation output.

3 Experiments and Results

For both MS T2 lesion segmentation and brain tumour segmentation pipelines, we investigate the network's performance in 3 settings: (1) Task-2 networks are trained on the same input MRI as Task-1, (2) In addition to MRI, the Task-2 network also takes the deterministic prediction of the Task-1 network as input, (3) MRI, deterministic predictions, and the uncertainty of the Task-1 network output are provided as input to the Task-2 network.

3.1 MS T2 Lesion Segmentation/Detection Pipeline

In this pipeline, both tasks consist of T2 lesion segmentation networks. For the first task, we train BU-Net [5] (Task-1 network) to segment T2 lesions given



(A) (B)

Fig. 2. Receiver-operating characteristic (ROC) curves comparing overall MS T2 lesion detection performance, illustrating TPR (true positive rate) vs. FDR (false detection rate) across (A) all lesions, and (B) small lesions (3–10 voxels) with several input configurations. At the operating point of FDR = 0.2, TPR values are (A) 0.44, 0.47, and 0.49 and (B) 0.78, 0.80, and 0.84 for MR sequence, MR sequence + segmentation, and MR sequence + segmentation + uncertainty experiments, respectively.

multiple MRI (T1, T2, FLAIR, and proton density (PD)), and provide a corresponding segmentation uncertainty. 10 Monte Carlo samples are used to compute the segmentation uncertainty. The second task consists of a modified 3D U-Net (Task-2 network) that again performs binary voxel-level T2 lesion segmentation. These voxel-level segmentations are subsequently grouped into discrete lesion-level instances.

A proprietary dataset, used for training and testing, consists of three multi-site, multi-scanner clinical relapsing-remitting MS (RRMS) trials, with a total of 2832 patients at different disease stages, resulting in over 5800 multi-modal MRI (T1, T2, FLAIR, and PD). The majority of these patients were scanned annually or bi-annually over a 24-month period. MRI sequences were acquired at $1\,\text{mm} \times 1\,\text{mm} \times 3\,\text{mm}$ resolution. Expert T2 lesion labels provided with the dataset were the result of expert human annotators manually correcting an automated segmentation method. 40% of the available data was used for training/validating the Task-1 network, with a 90/10 training/validation split. 50% of the available data was used for training/validating the Task-2 network, again with a 90/10 training/validation split. Finally, 10% of the available data is used for testing the Task-2 network. The dataset is divided this way in order to provide the Task-2 network with consistent and meaningful uncertainties.

Since the downstream outcome of interest is the accurate detection of T2 lesions, we evaluate the performance of Task-1 and Task-2 networks based on lesion-level TPR and FDR. To obtain lesion-level detections from the voxel-based segmentations provided by the Task-2 network, a connected component analysis is performed to group lesion voxels together in an 18-connected neighbourhood [5]. The TPR and FDR are then calculated at the lesion level and are used to plot receiver operating characteristic (ROC) curves. Given that MS lesions

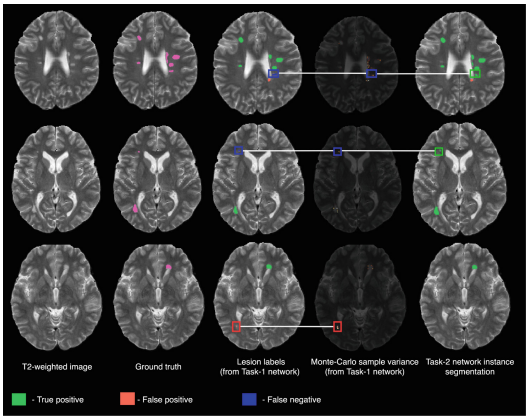


Fig. 3. Examples demonstrating the corrective effect of uncertainty propagation on MS lesion detection performance for three patients. Images are of, from left to right, T2 weighted MRI input, ground truth T2 lesion labels (in magenta), T2 lesion labels produced by the Task-1 network, the MC sample variance uncertainty estimates at the Task-1 network output, and the T2 lesion labels produced by the Task-2 network.

vary greatly in size, the system performance is evaluated on lesions grouped into three size bins: small (3–10 vox), medium (11–50 vox), and large (51+ vox). The system performs almost perfectly in detecting medium and large lesions in all 3 settings. Given that the detection of small lesions is particularly challenging, and that 40% of the lesions in the dataset are small, lesion detection ROC results are reported for both the overall performance on all lesions and then on only the small lesions in Fig. 2. At an FDR of 0.2 (the clinical operating point of interest), the TPR performance increases by 4% for small lesions and 5% for all lesions with the inclusion of the Task-1 uncertainty map as an additional input to the Task-2 network. Qualitative results for three cases are provided in Fig. 3 illustrate how the propagation of the uncertainty information enables the correction of both false positives (bottom case) and false negatives (top two cases).

3.2 Brain Tumour Segmentation Pipeline

Segmentation of brain tumours requires the presence of multiple MRI sequences (T1, T2, T1ce, and FLAIR) that provide complementary information. In clinical scenarios, one or more of these critical MR sequences might be missing (e.g. FLAIR), due to a variety of reasons, including cost or time constraints, noise in acquisition, etc. One way to address this is to synthesize the missing sequence before tumour segmentation [2,11]. We train RS-Net (Task-1 network) [2] to synthesize T1ce and FLAIR MRI, and generate its corresponding synthesis (regression) uncertainty. We synthesize these sequences as previous work [2,11] has shown that their absence will decrease segmentation performance more than the absence of either T1 or T2. T1ce is the hardest sequence to synthesize because

Table 1. Comparison of multi-class brain tumour segmentation based on modified 3D U-Net on the BraTS 2018 Validation dataset. The inclusion of the associated uncertainties from Task-1, in addition to the RS-Net synthesis output, as input to the 3D U-Net network is shown to lead to improvements on the Dice values. Quantitative segmentation results are based on percentage Dice coefficients for: enhancing tumor (DE), whole tumor (DT), and tumor core (DC). (*) indicates statistically significant ($p \leq 0.05$) differences between second and third row.

	T1ce synthesis			FLR synthesis		
	DT	DC	DE	DT	DC	DE
real(3) sequences	87.17	50.25	26.89	83.27	73.91	71.07
real(3)+synthesized sequences	86.72	52.80	27.35	84.56	76.72	72.89
real(3)+synthesized+uncertainty	**88.20**	**57.29***	**32.86***	**85.84***	**79.25***	**74.51***

it shows enhancement within the tumour resulting from a contrast agent, which is not present in the other MRI sequences used in its synthesis. RS-Net uses T1, T2, and FLAIR to synthesize T1ce, and T1, T1ce, and T2 to synthesize FLAIR. Uncertainties associated with synthesized MRI are estimated using 20 MC samples. We then train a modified 3D U-Net (Task-2 network) for multi-class brain tumour segmentation, comparing three experiment settings detailed in Sect. 3.

This pipeline is evaluated using the 2018 MICCAI BraTS [13] dataset. The BraTS training dataset is comprised of 210 HGG and 75 LGG patients with different MRI sequences: T1, T1ce, T2, and FLAIR MRI for each patient. Ground truth tumour labels were provided by expert human annotators, and consist of 3 classes: edema, necrotic/non-enhancing core, and enhancing tumor core. 228 patients were randomly selected for training the network and another remaining 57 for network validation. A separate BraTS 2018 validation dataset was used to test the segmentation performance. This dataset contains 66 patient multi-channel MRI (with no labels provided). The BraTS challenge provides pre-processed volumes that were skull-stripped, co-aligned, and resampled to $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ voxel resolution. The intensities were additionally normalized using mean subtraction, divided by the standard deviation, and rescaled from 0 to 1, using the brain-masked region of a given MR image. The images were then cropped to $184 \times 200 \times 152$. To make sure that regression uncertainties are associated with data that was not seen during training, the RS-Net was trained in two folds, with each fold comprised of 114 volumes. The segmentation U-Net was trained using all 228 volumes in a single fold.

The performance of the brain tumour segmentation is evaluated by calculating Dice scores for three different tumour sub-types: enhancing tumour, whole tumour, and tumour core. This is consistent with the evaluation metrics used in the BraTS challenge [13]. Quantitative results (Table 1) indicate that when the associated regression uncertainty is propagated as an input to the segmentation (3D U-Net) network in addition to the synthesized MRI, the network performance increases by either 2–10% (T1ce synthesis) or 2–5% (FLAIR synthesis), over propagating only the synthesized MRI. As seen in other works [2,11] and

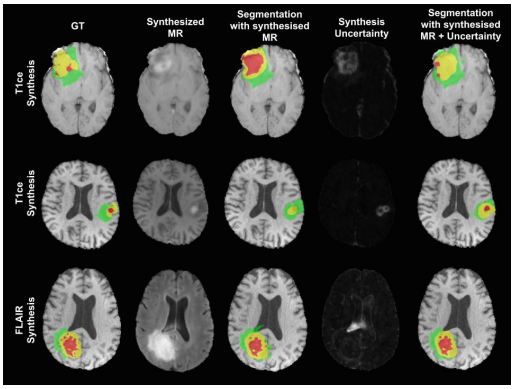


Fig. 4. Examples demonstrating the 3D U-Net performance on the multi-class brain tumour segmentation task based on synthesized MRI sequences. From Left to Right: GT segmentation, synthesized MR sequence, segmentation using real MRI (3 sequences) + synthesized MRI, synthesis uncertainty, segmentation using real MRI (3 sequences) + synthesized MRI + synthesis uncertainty. First two rows: T1ce synthesis. Last row: FLAIR synthesis. Labels: edema (green), non-enhancing or necrotic tumour core (red), enhancing tumour (yellow). (Color figure online)

mentioned above, the overall network performance is lower for T1ce as compared to FLAIR as it is more challenging to synthesize.

Figure 4 shows visual examples of the results on the downstream brain tumour segmentation when MRI sequences are synthesized. In the first row, it is clear that the framework that does not include synthesis uncertainty results in confusion between enhancing tumour and core tumour, as the enhancement is not well captured in the synthesized T1ce. However, the synthesis uncertainty is higher in this region. Consequently, propagating the uncertainty information informs the Task-2 (segmentation) network about mistakes made by Task-1 (synthesis) network, thereby enabling the correction of these errors. Similarly, in the second row, we can see that uncertainty provides supplementary information to the synthesized T1ce and allows the network to correctly identify enhancing and non-enhancing core. In the third row, the FLAIR sequence is synthesized. Here, the network incorrectly segments background near and within the ventricle as edema when uncertainty is not propagated. This is because the ventricle is incorrectly highlighted in this area in the synthesized FLAIR. As the uncertainty for this synthesized region is high, cascading the uncertainties permits the network to learn to correct its error.

4 Conclusions

This work proposes a general deep learning framework for the propagation of uncertainty across a sequence of inference tasks within a medical image analysis pipeline, and demonstrated that cascading uncertainties (e.g. based on MC

dropout) in this manner can lead to improvements in performance for the downstream task. The framework was applied to two different contexts. First, it was demonstrated that by propagating voxel-based lesion segmentation uncertainties to a second segmentation network, lesion-level detection performance can be improved (in terms of a reduction of both FPs and FNs) based on experiments on a large-scale, multi-site, clinical dataset of patients with MS. Next, it was also demonstrated that by propagating regression uncertainty from an MRI synthesis task, performance of a downstream multi-class tumour segmentation task can be improved based on experiments on the publicly available BraTS dataset. Future work will explore how to properly develop a complete end-to-end system that includes uncertainty propagation across the inference modules.

Acknowledgements. This work was supported by a Canadian Natural Science and Engineering Research Council (NSERC) Collaborative Research and Development Grant (CRDPJ 505357 - 16), Synaptive Medical, the Canadian NSERC Discovery and CREATE grants, and an award from the International Progressive MS Alliance (PA-1603-08175).

References

1. Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A.: Multimodal MR synthesis via modality-invariant latent representation. IEEE Trans. Med. Imaging **37**(3), 803–814 (2017)
2. Mehta, R., Arbel, T.: RS-Net: regression-segmentation 3D CNN for synthesis of full resolution missing brain MRI in the presence of tumours. In: Gooya, A., Goksel, O., Oguz, I., Burgos, N. (eds.) SASHIMI 2018. LNCS, pp. 119–129. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00536-8_13
3. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 729–738. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_82
4. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21
5. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 655–663. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_74
6. Tousignant, A., Lemaître, P., Precup, D., Arnold, D.L., Arbel, T.: Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data. In: International Conference on Medical Imaging with Deep Learning, pp. 483–492, May 2019
7. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059, June 2016

8. Ozdemir, O., Woodward, B., Berlin, A.A.: Propagating uncertainty in multi-stage Bayesian convolutional neural networks with application to pulmonary nodule detection. arXiv preprint arXiv:1712.00497 (2017)
9. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Alzheimer's Disease Neuroimaging Initiative: Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage **195**, 11–22 (2019)
10. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. **7**(1), 17816 (2017)
11. van Tulder, G., de Bruijne, M.: Why does synthesized data improve multi-sequence classification? In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 531–538. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_65
12. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
13. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)
14. Wu, Y., He, K.: Group normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_1
15. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)