Generalized Wasserstein Dice Loss, Test-Time Augmentation, and Transformers for the BraTS 2021 Challenge

Lucas Fidon[1(✉)], Suprosanna Shit[2(✉)], Ivan Ezhov[2], Johannes C. Paetzold[2], Sébastien Ourselin[1], and Tom Vercauteren[1]

[1] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
lucas.fidon@kcl.ac.uk

[2] Department of Informatics, Technical University of Munich, Munich, Germany
suprosanna.shit@tum.de

Abstract. Brain tumor segmentation from multiple Magnetic Resonance Imaging (MRI) modalities is a challenging task in medical image computation. The main challenges lie in the generalizability to a variety of scanners and imaging protocols. In this paper, we explore strategies to increase model robustness without increasing inference time. Towards this aim, we explore finding a robust ensemble from models trained using different losses, optimizers, and train-validation data split. Importantly, we explore the inclusion of a transformer in the bottleneck of the U-Net architecture. While we find transformer in the bottleneck performs slightly worse than the baseline U-Net in average, the generalized Wasserstein Dice loss consistently produces superior results. Further, we adopt an efficient test time augmentation strategy for faster and robust inference. Our final ensemble of seven 3D U-Nets with test-time augmentation produces an average dice score of 89.4% and an average Hausdorff 95% distance of 10.0 mm when evaluated on the BraTS 2021 testing dataset. Our code and trained models are publicly available at https://github.com/LucasFidon/TRABIT_BraTS2021.

Keywords: BraTS 2021 · Segmentation · Deep learning · Brain tumor · Transformers · Test-time augmentation

1 Introduction

Gliomas are the most common malignant brain tumors. Broadly, Gliomas are categorized into aggressive high-grade and slow-growing low-grade types. In both types of Gliomas, changes in tissues caused by tumor cells can be captured using multi-modality Magnetic Resonance Imaging (MRI). The commonly used modalities are T1, T2, contrast-enhanced T1 (ceT1), and FLAIR. These modalities are

A. Crimi and S. Bakas (Eds.): BrainLes 2021, LNCS 12963, pp. 187–196, 2022.
https://doi.org/10.1007/978-3-031-09002-8_17

the default choice for the radiologist to identify the tumor type and its progression stage. Towards this objective, accurate and automatic brain tumor segmentation based on multi-parametric MRI is an active field of research [23] and could support diagnosis, surgery planning [11,12], follow-up, and radiation therapy [1,2]. The BraTS 2021 challenge has offered an unique and unprecedented opportunity to machine learning researchers to develop a clinically deployable solution for Glioma multi-class segmentation.

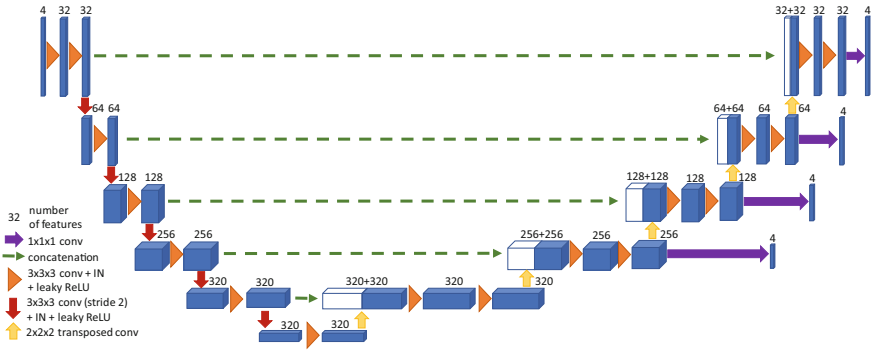

Fig. 1. Illustration of the 3D U-Net [10] **architecture used.** Blue boxes represent feature maps. IN stands for instance normalization [32]. The design of this 3D U-Net was determined using the heuristics of nnU-Net and our previous work [13,16,17,21]. (Color figure online)

Aiming for computational efficiency, we use 3D U-Net, and its recent transformer variation, TransUNet [9], as the primary models and focus on finding better learning schemes, such as, augmentation, loss function, optimizer, and efficient inference routine for ensemble model. Recently it has been shown that different loss function combinations may have a crucial impact on the resultant segmentation [24]. In our settings, we use the Generalized Wasserstein Dice loss [15] that has shown superior segmentation performance as compared to the mean Dice loss [26,28,30] in the BraTS 2020 challenge [16] and for other medical image segmentation tasks [8,31]. We investigate the effect of different state-of-the-art optimizers, such as, SGD, SGDP [20], ASAM [25]. Lastly, we use an efficient test-time ensemble approach for the final segmentation result.

2 Methods and Materials

2.1 Data

We have used the BraTS 2021 dataset[1] [3] in our experiments. No additional data were used. The dataset contains the same four MRI sequences (T1, ceT1,

[1] https://www.synapse.org/#!Synapse:syn25829067/wiki/610865.

T2, and FLAIR) for all cases, corresponding to patients with either a high-grade Gliomas [5] or a low-grade Gliomas [6]. All the cases were manually segmented for peritumoral edema, enhancing tumor, and non-enhancing tumor core using the same labeling protocol [3,4,7,27]. The training dataset contains 1251 cases, and the validation dataset contains 219 cases. MRI for training and validation datasets are publicly available, but only the manual segmentations for the training dataset are available. The evaluation on the validation dataset was performed using the BraTS 2021 challenge online evaluation platform[2]. For each case, the four MRI sequences are available after co-registration to the same anatomical template, interpolation to 1mm isotropic resolution, and skull stripping [27].

2.2 Deep Learning Pipeline

We used the DynU-Net of MONAI [29] to implement a baseline 3D U-Net with one input block, 4 down-sampling blocks, one bottleneck block, 5 upsampling blocks, 32 features in the first level, instance normalization [32], and leaky-ReLU with slope 0.01. An illustration of the architecture is provided in Fig. 1. We have used the same pipeline for our participation to the FeTA challenge 2021 [14].

Transformers have recently received attention in medical image computing for their multi-hop attention mechanism. As a second network architecture, we replace the bottleneck block of the U-Net with a vision transformer as proposed by [9]. We use the identical transformer architecture for our experiment as in [9]. A transformer in the bottleneck allows to accumulate the global context of the image and learn an anatomically consistent representation of the tumor classes.

Table 1. Network architecture specification

Network	No. of parameter	Avg. inference time
3D U-Net [10]	31.2M	6 s
TransUNet [9]	116.7M	10 s

Table 1 shows a comparison in terms of the number of parameters and inference time between 3D U-Net and transUNet. For both networks, we train using a patch size of $128 \times 192 \times 128$.

2.3 Loss Function

We have experimented with two loss functions: the sum of the cross-entropy loss and the mean-class Dice loss

$$\mathcal{L}_{DL+CE} = \mathcal{L}_{DL} + \mathcal{L}_{CE} \tag{1}$$

[2] https://www.synapse.org/#!Synapse:syn25829067/wiki/.

and the sum of the cross entropy loss and of the generalized Wasserstein Dice loss[3] [15,16].

$$\mathcal{L}_{GWDL+CE} = \mathcal{L}_{GWDL} + \mathcal{L}_{CE} \tag{2}$$

where $\mathcal{L}_{CE}$ is the cross entropy loss function

$$\mathcal{L}_{CE}(\hat{\mathbf{p}}, \mathbf{p}) = -\sum_{i=1}^{N}\sum_{l=1}^{L} p_{i,l}\log(\hat{p}_{i,l}) \tag{3}$$

with N the number of voxels, L the number of classes, i the index for voxels, l the index for classes, $\hat{\mathbf{p}} = (\hat{p}_{i,l})_{i,l}$ the predicted probability map, and $\mathbf{p} = (p_{i,l})_{i,l}$ the discrete ground-truth probability map.

$\mathcal{L}_{DL}$ is the mean-class Dice loss [26,30]

$$\mathcal{L}_{DL}(\hat{\mathbf{p}}, \mathbf{p}) = 1 - \frac{1}{L}\sum_{l=1}^{L}\frac{2\sum_{i=1}^{N} p_{i,l}\hat{p}_{i,l}}{\sum_{i=1}^{N} p_{i,l} + \sum_{i=1}^{N}\hat{p}_{i,l}} \tag{4}$$

And $\mathcal{L}_{GWDL}$ is the generalized Wasserstein Dice loss [15]

$$\begin{cases} \mathcal{L}_{GWDL}(\hat{\mathbf{p}}, \mathbf{p}) = 1 - \dfrac{2\sum_{l\neq b}\sum_i \mathbf{p}_{i,l}(1 - W^M(\hat{\mathbf{p}}_i, \mathbf{p}_i))}{2\sum_{l\neq b}[\sum_i p_{i,l}(1 - W^M(\hat{\mathbf{p}}_i, \mathbf{p}_i))] + \sum_i W^M(\hat{\mathbf{p}}_i, \mathbf{p}_i)} \\ \forall i, \quad W^M(\hat{\mathbf{p}}_i, \mathbf{p}_i) = \sum_{l=1}^{L} p_{i,l}\left(\sum_{l'=1}^{L} M_{l,l'}\hat{p}_{i,l'}\right) \end{cases} \tag{5}$$

where $W^M(\hat{\mathbf{p}}_i, \mathbf{p}_i)$ is the Wasserstein distance between predicted $\hat{\mathbf{p}}_i$ and ground truth $\mathbf{p}_i$ discrete probability distribution at voxel i. $M = (M_{l,l'})_{1\leq l,\, l'\leq L}$ is a distances matrix between the BraTS 2021 labels, and b is the class number corresponding to the background. For the classes indices 0: *background*, 1: *enhancing tumor*, 2: *edema*, 3: *non-enhancing tumor*, we set

$$M = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0.7 & 0.5 \\ 1 & 0.7 & 0 & 0.6 \\ 1 & 0.5 & 0.6 & 0 \end{pmatrix} \tag{6}$$

The generalized Wasserstein Dice loss [15] is a generalization of the Dice Loss for multi-class segmentation that can take advantage of the hierarchical structure of the set of classes in BraTS. When the labeling of a voxel is ambiguous or too difficult for the neural network to predict it correctly, the generalized Wasserstein Dice loss and our matrix M are designed to favor mistakes that remain consistent with the sub-regions used in the evaluation of BraTS, i.e., core tumor and whole tumor.

[3] https://github.com/LucasFidon/GeneralizedWassersteinDiceLoss.

2.4 Optimization

Common Optimization Setting: For each network, the training dataset was split into 95% training and 5% validation at random. The random initialization of the weights was performed using He initialization [19] for all the deep neural network architectures. We used batch size 2. The CNN parameters used at inference corresponds to the last epoch. We used deep supervision with 4 levels during training. Training each 3D U-Net required 16GB of GPU memory.

SGD: SGD with Nesterov momentum. The initial learning rate was 0.02, and we used polynomial learning rate decay with power 0.9 for a total of 500 epochs.

ADAM [22]**:** For Adam, we used a linear warmup for 1000 iterations for the learning rate from 0 to 0.003 followed by a constant learning rate schedule at the value 0.003 for 500 epochs.

Adaptive Sharpness-Aware Minimization (ASAM) [18,25]**:** We have used SGD as the base optimizer with the initial learning rate set to 0.02 and we used polynomial learning rate decay with power 0.9 for a total of 500 epochs. We used the default hyperparameters of ASAM [25], $\rho = 0.5$, and $\eta = 0.1$. We have used the PyTorch implementation of the authors[4].

SGDP [20]**:** For SGD Projected (SGDP), we have used the exact same hyperparameter values as for SGD. We have used the PyTorch implementation of the authors[5].

2.5 Data Augmentation

We have used random zoom (zoom ratio range $[0.7, 1.5]$ drawn uniformly at random; probability of augmentation 0.3), random rotation (rotation angle range $[-15°, 15°]$ for all dimensions drawn uniformly at random; probability of augmentation 0.3), random additive Gaussian noise (mean 0, standard deviation 0.1; probability of augmentation 0.3), random Gaussian spatial smoothing (standard deviation range $[0.5, 1.5]$ in voxels for all dimensions drawn uniformly at random; probability of augmentation 0.2), random gamma augmentation (gamma range $[0.7, 1.5]$ drawn uniformly at random; probability of augmentation 0.3), and random right/left flip (probability of augmentation 0.5).

2.6 Inference

Single Models Inference: For the models evaluated and compared in Fig. 2, a patch-based approach is used. The input image is divided into overlapping patches of size $128 \times 192 \times 128$. The patches are chosen, so that neighboring patches have an overlap of at least half of their volume. The fusion of the patch

[4] https://github.com/SamsungLabs/ASAM.
[5] https://github.com/clovaai/AdamP.

prediction is performed using a weighted average of the patch predictions before the softmax operation. The weights are defined with respect to the distance of a voxel to the center of the patch using a Gaussian kernel standard deviation for each dimension equal to $0.125 \times$ patch-dimension. In addition, test-time augmentation [33] is used with right-left flip. The two softmax predictions obtained with and without right-left flip are merged by averaging.

Ensemble Inference: For the ensembles, the inference is performed in two steps. During the first step, a first segmentation is computed using only one model and the inference procedure for single models. In practice, we used the first model of the list and did not tune the choice of this model.

The first segmentation is used to estimate the center of gravity of the whole tumor. In the second step, we crop a patch of size $128 \times 192 \times 128$ with a center chosen as close as possible to the center of gravity of the tumor so that the patch fits in the image. The segmentation probability predictions of all the models of the ensemble are computed for this patch. The motivation for this two-step approach is to reduce the inference time as compared to using the patch-based approach described above for all the models of the ensemble. This strategy is based on the assumption that a patch of size $128 \times 192 \times 128$ is large enough to always contain the whole tumor. During the second step, test-time augmentations with right-left flip and zoom with a ratio of 1.125 are used. The four segmentation probability predictions obtained for the different augmentations (no flip - no zoom, flip - no zoom, no flip - zoom, and flip - zoom) are combined by averaging the softmax predictions. For the full image segmentation prediction, the voxels outside the patch centered on the tumor are set to the background.

3 Results

As a primary metric, we report the mean and the standard deviation of the Dice score and the Hausdorff distance for each class. Percentiles are common statistics for measuring the robustness of automatic segmentations [13]. To evaluate the robustness of the different models, we report the percentiles of the Dice score at 25% and 5% and the percentiles at 75% and 95% of the Hausdorff 95% distance. In Table 2, we report the validation scores of our individually trained models. In Table 3 we compare two ensemble strategies as described in the previous section. In the ensemble models, we don't include the TransUNet model as their individual performance is marginally worse than the 3D U-Net model.

4 Discussion

From Table 2, we see that 3D U-Net trained with generalized Wasserstein Dice loss performs consistently better than the one with Dice loss (baseline model). TransUNet does not offer any improvement over the baseline. Rather the performance deteriorates slightly. We hypothesize that over-parameterization can

Table 2. Segmentation results on the BraTS 2021 Validation dataset. The evaluation was performed on the BraTS online evaluation platform. ET: Enhancing Tumor, WT: Whole Tumor, TC: Tumor Core, Std: Standard deviation, p_x: Percentile x. The split number corresponds to the random seed that was used to split the training dataset into 95% training/5% validation at random.

		Dice score (%)				Hausdorff 95% (mm)			
Model	ROI	Mean	Std	p_{25}	p_5	Mean	Std	p_{75}	p_{95}
3D U-Net	ET	82.6	23.7	83.6	7.7	17.9	73.7	2.2	25.4
GWDL + CE SGD	TC	86.4	20.1	86.8	37.7	11.2	50.1	4.2	19.4
Split 1	WT	92.5	7.4	90.7	82.1	3.8	5.7	3.6	17.7
3D U-Net	ET	82.2	24.0	82.9	7.2	17.8	73.7	2.4	18.1
GWDL + CE SGD	TC	86.5	20.4	86.1	37.7	11.1	50.1	4.4	21.3
Split 2	WT	92.5	7.4	90.6	82.6	3.8	5.9	3.7	11.5
3D U-Net	ET	81.9	24.5	82.5	5.3	19.5	77.5	2.2	31.7
GWDL + CE SGD	TC	85.7	21.6	86.0	30.5	11.5	50.1	4.2	20.8
Split 27	WT	92.5	7.1	90.5	81.7	4.0	6.2	3.9	10.6
3D U-Net	ET	81.9	24.9	83.0	0.0	21.1	81.0	2.2	67.0
GWDL + CE SGD	TC	86.5	20.7	87.2	38.5	9.5	43.7	4.1	16.9
Split 1227	WT	92.5	7.3	90.6	81.1	3.8	5.8	3.7	11.7
3D U-Net	ET	82.4	24.4	83.2	2.9	19.5	77.5	2.4	30.3
GWDL + CE SGD	TC	85.8	21.7	86.3	27.6	11.5	50.1	4.6	23.0
Split 122712	WT	92.4	7.1	90.2	81.7	4.1	7.0	3.9	11.4
3D U-Net	ET	81.7	24.9	82.9	0.0	21.3	81.1	2.4	89.3
DL + CE SGD	TC	86.5	20.4	86.3	40.2	11.0	50.1	4.1	17.9
Split 1	WT	92.5	7.2	90.6	80.2	3.9	6.7	3.9	9.8
3D U-Net	ET	82.1	24.2	82.8	5.4	19.5	77.5	2.5	31.3
GWDL + CE SGDP	TC	86.3	20.5	86.4	40.4	9.7	43.7	4.1	20.2
Split 1	WT	92.6	7.4	90.5	82.0	3.8	5.9	3.7	11.0
3D U-Net	ET	80.0	25.7	81.5	0.0	23.4	84.5	3.0	373.1
GWDL + CE ASAM	TC	85.4	21.5	86.1	35.3	12.2	50.8	4.2	21.7
Split 1	WT	91.9	7.9	90.0	77.2	5.4	10.7	4.2	22.0
TransUNet	ET	79.7	25.3	80.1	0.0	22.4	81.0	3.0	113.4
GWDL + CE SGD	TC	84.5	22.0	84.5	36.5	8.7	36.5	4.9	24.9
Split 1	WT	91.8	7.3	90.2	77.2	4.4	7.6	4.1	12.7
TransUNet	ET	80.9	24.1	81.2	6.5	18.5	73.6	3.0	37.7
GWDL + CE ADAM	TC	84.8	21.9	84.7	36.44	10.3	43.8	4.6	24.6
Split 1	WT	91.9	7.8	90.0	80.1	4.1	6.5	4.1	14.9
TransUNet	ET	80.2	25.5	80.4	0.0	23.1	84.3	3.0	373.1
GWDL + CE SGDP	TC	85.2	21.9	85.6	27.6	11.45	50.0	4.5	19.9
Split 1	WT	92.0	7.4	90.1	79.7	4.7	9.1	4.0	18.0

Table 3. Segmentation results on the BraTS 2021 Validation dataset for ensembling and test-time augmentation. The evaluation was performed on the BraTS online evaluation platform. ET: Enhancing Tumor, WT: Whole Tumor, TC: Tumor Core, Std: Standard deviation, p_x: Percentile x. Best values are in **bold**.

		Dice score (%)				Hausdorff 95% (mm)			
Model	ROI	Mean	Std	p_{25}	p_5	Mean	Std	p_{75}	p_{95}
3D U-Net	ET	82.0	24.4	83.0	3.1	19.5	77.5	2.3	30.0
Ensemble	TC	86.6	20.2	86.5	39.2	9.5	43.7	4.1	20.1
	WT	92.6	7.2	90.6	82.0	3.9	6.3	3.6	12.8
3D U-Net	ET	**84.0**	**22.0**	**84.2**	**21.6**	**12.7**	**60.6**	**2.2**	**16.0**
Ensemble	TC	**87.0**	**20.0**	**86.8**	**43.1**	**11.0**	**50.1**	**4.1**	**18.7**
Zoom augmentation	WT	**92.7**	**7.2**	**90.6**	**82.1**	**3.9**	**6.3**	**3.6**	**12.6**

Table 4. Segmentation results on the BraTS 2021 Testing dataset using ensembling and test-time augmentation. The evaluation was performed by the BraTS 2021 challenge organizers using our docker submission. ET: Enhancing Tumor, WT: Whole Tumor, TC: Tumor Core, Std: Standard deviation, p_x: Percentile x.

		Dice score (%)			Hausdorff 95% (mm)		
Model	ROI	Mean	Std	p_{25}	Mean	Std	p_{75}
3D U-Net	ET	87.4	17.6	85.2	10.1	53.5	2.0
Ensemble	TC	87.8	23.6	91.3	15.8	66.7	3.0
Zoom augmentation	WT	92.9	9.0	91.6	4.1	7.3	3.7

be an issue in this case. The optimizer SGDP and ASAM perform similar to the baseline SGD. From Table 3, we see that ensemble strategy helps in increasing the robustness of the model. The best ensemble strategy turns out to be including zoom as a test time augmentation. This approach was submitted for evaluation on the BraTS 2021 testing dataset and the results can be found in Table 4. In conclusion, this paper proposes a detailed comparative study on the strategies to make a computationally efficient yet robust automatic brain tumor segmentation model. We have explored ensemble from multiple training configurations of different state-of-the-art loss functions and optimizers, and importantly, test-time augmentation. Future research will focus on further strategies on test-time augmentation and test-time hyper-parameter tuning.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement TRABIT No 765148; Wellcome [203148/Z/16/Z; WT101957], EPSRC [NS/A000049/1; NS/A000027/1]. Tom Vercauteren is supported by a Medtronic/RAEng Research Chair [RCSRF1819\7\34]. Data used in this publication were obtained as part of the RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge project through Synapse ID (syn25829067).

References

1. Andres, E.A., et al.: Po-1002 pseudo computed tomography generation using 3D deep learning-application to brain radiotherapy. Radiother. Oncol. **133**, S553 (2019)
2. Andres, E.A., et al.: Dosimetry-driven quality measure of brain pseudo computed tomography generated from deep learning for MRI-only radiotherapy treatment planning. Int. J. Radiat. Oncol.* Biol.* Phys. **108**, 813–823 (2020)
3. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
4. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
5. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. Cancer Imaging Arch. (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
6. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Cancer Imaging Arch. (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
7. 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)
8. Blanc-Durand, P., et al.: Prognostic value of anthropometric measures extracted from whole-body CT using deep learning in patients with non-small-cell lung cancer. Eur. Radiol. **30**(6), 3528–3537 (2020). https://doi.org/10.1007/s00330-019-06630-w
9. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
10. 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
11. Ezhov, I., et al.: Neural parameters estimation for brain tumor growth modeling. In: MICCAI 2019. LNCS, vol. 11765, pp. 787–795. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_87
12. Ezhov, I., et al.: Geometry-aware neural solver for fast Bayesian calibration of brain tumor models. arXiv preprint arXiv:2009.04240 (2020)
13. Fidon, L., et al.: Distributionally robust segmentation of abnormal fetal brain 3D MRI. In: UNSURE/PIPPI -2021. LNCS, vol. 12959, pp. 263–273. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87735-4_25
14. Fidon, L., et al.: Partial supervision for the feta challenge 2021. arXiv preprint arXiv:2111.02408 (2021)
15. Fidon, L., et al.: Generalised Wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 64–76. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_6
16. Fidon, L., Ourselin, S., Vercauteren, T.: Generalized Wasserstein dice score, distributionally robust deep learning, and ranger for brain tumor segmentation: Brats 2020 challenge. arXiv preprint arXiv:2011.01614 (2020)

17. Fidon, L., Ourselin, S., Vercauteren, T.: SGD with hardness weighted sampling for distributionally robust deep learning. arXiv preprint arXiv:2001.02658 (2020)
18. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)
19. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
20. Heo, B., et al.: AdamP: slowing down the slowdown for momentum optimizers on scale-invariant weights. arXiv preprint arXiv:2006.08217 (2020)
21. Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 (2020)
22. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Kofler, F., Berger, C., Waldmannstetter, D., Lipkova, J., Ezhov, I., Tetteh, G., Kirschke, J., Zimmer, C., Wiestler, B., Menze, B.H.: Brats toolkit: translating brats brain tumor segmentation algorithms into clinical and scientific practice. Front. Neurosci. **14**, 125 (2020)
24. Kofler, F., et al.: Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the dice coefficient. arXiv preprint arXiv:2103.06205 (2021)
25. Kwon, J., Kim, J., Park, H., Choi, I.K.: ASAM: adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. arXiv preprint arXiv:2102.11600 (2021)
26. Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T.: On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 348–360. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_28
27. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
28. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
29. MONAI Consortium: MONAI: Medical open network for AI, March 2020. https://doi.org/10.5281/zenodo.4323058. https://github.com/Project-MONAI/MONAI
30. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
31. Tilborghs, S., et al.: Comparative study of deep learning methods for the automatic segmentation of lung, lesion and lesion type in CT scans of COVID-19 patients. arXiv preprint arXiv:2007.15546 (2020)
32. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
33. 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 (2019)