MVP U-Net: Multi-View Pointwise U-Net for Brain Tumor Segmentation

Changchen Zhao, Zhiming Zhao, Qingrun Zeng, and Yuanjing Feng(✉)

Zhengjiang University of Technology, Hangzhou 310023, China
fyjing@zjut.edu.cn

Abstract. It is a challenging task to segment brain tumors from multi-modality MRI scans. How to segment and reconstruct brain tumors more accurately and faster remains an open question. The key is to effectively model spatial-temporal information that resides in the input volumetric data. In this paper, we propose Multi-View Pointwise U-Net (MVP U-Net) for brain tumor segmentation. Our segmentation approach follows encoder-decoder based 3D U-Net architecture, among which, the 3D convolution is replaced by three 2D multi-view convolutions in three orthogonal views (axial, sagittal, coronal) of the input data to learn spatial features and one pointwise convolution to learn channel features. Further, we modify the Squeeze-and-Excitation (SE) block properly and introduce it into our original MVP U-Net after the concatenation section. In this way, the generalization ability of the model can be improved while the number of parameters can be reduced. In BraTS 2020 testing dataset, the mean Dice scores of the proposed method were 0.715, 0.839, and 0.768 for enhanced tumor, whole tumor, and tumor core, respectively. The results show the effectiveness of the proposed MVP U-Net with the SE block for multi-modal brain tumor segmentation.

Keywords: Multi-View Pointwise U-Net · Brain tumor segmentation · BraTS 2020 · Spatial-temporal network · SE block

1 Introduction

Qualitative and quantitative assessment of brain tumors is the key to determine whether medical images can be used in clinical diagnosis and treatment. Researchers began to explore faster and more accurate methods for brain tumor segmentation. However, due to the fuzziness of the boundaries of each tumor subregion, the complete automatic segmentation of brain tumors remains challenging.

Brain Tumor Segmentation (BraTS) Challenge [1–4,17] has always been focusing on the evaluation of state-of-the-art methods for the segmentation of brain tumors in multimodal magnetic resonance imaging (MRI) scans. BraTS 2020 utilizes multi-institutional pre-operative MRI scans and primarily focuses on the segmentation of intrinsically heterogeneous brain tumors, namely gliomas.

A. Crimi and S. Bakas (Eds.): BrainLes 2020, LNCS 12659, pp. 93–103, 2021.
https://doi.org/10.1007/978-3-030-72087-2_9

The BraTS 2020 dataset is annotated manually by one to four raters, following the same annotation protocol, and their annotations are approved by experienced neuro-radiologists. Annotations comprise the background (label 0), the enhancing tumor (ET, label 4), the peritumoral edema (ED, label 2), and the necrotic and non-enhancing tumor (NCR/NET, label 1). Each patient's MRI scan consists of four modalities, i.e., native T1 weighted, post-contrast T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR).

Since the U-Net network was first proposed by Ronneberger et al. in 2015 [18], the neural network represented by U-Net and its variants has been shining brightly in the field of medical image segmentation. It is a specialized convolutional neural network (CNN) with a down-sampling encoding path and an up-sampling decoding path similar to an auto-encoder architecture. However, because the U-Net network takes as input two-dimensional data while medical images are usually three-dimensional, using the U-Net network will lose the spatial details of the original data. As a result, the image segmentation accuracy is not satisfying. Alternatively, 3D U-Net [8] was proposed and has been widely used for segmentation in medical image segmentation due to its outstanding performance. However, 3D U-Net network is prone to overfitting and difficult to train because of its huge number of model parameters, which greatly limits its application. Both 2D and 3D U-Net models have their own advantages and disadvantages. One question naturally arises, is it possible to build a neural network that computes as low cost as a 2D network, but performs as well as a 3D network?

Researchers have been investigating this question for a long time and numerous approaches have been proposed. Haquer et al. [9] proposed 2.5D U-Net which consists of three 2D U-Net trained with axial, coronal, and sagittal slices, respectively. Although it achieves the goal of lower computation cost of 2D U-Net and the effectiveness of 3D U-Net, it does not make full use of the spatial information of volumetric medical image data. Chen et al. [6] proposed S3D U-Net which uses separable 3D convolutions instead of 3D convolutions. Although its segmentation of medical images is efficient, the number of model parameters is still large, which greatly limits its application in practical scenarios. It is a challenging task to achieve both low computational cost and high performance. The key is how to explore the spatial-temporal modeling for the input volumetric data. Recently, spatial-temporal 3D networks have received more and more attention [15]. It performs 2D convolution along three orthogonal views of volumetric video data to learn the spatial appearance and temporal motion cues, respectively, and fuses together to obtain the final output. Inspired by this, we propose Multi-View Pointwise U-Net (MVP U-Net) for brain tumor segmentation. The proposed MVP U-Net follows the encoder-decoder-based 3D U-Net architecture in which we use three multi-view convolutions and one pointwise convolution to reconstruct the 3D convolution.

Meanwhile, the Squeeze-and-Excitation (SE) block, proposed by Hu et al. [11] in 2017, can be incorporated into existing state-of-the-art deep CNN architecture such as ResNet and DenseNet as a subunit structure, and it can further improve the generalization ability of the original network by explicitly modeling the interdependencies between channels and adaptively calibrating the characteristic responses of channel correlation. In view of this, we incorporate this block into our MVP U-Net after appropriate modification.

2 Methods

2.1 Preprocessing

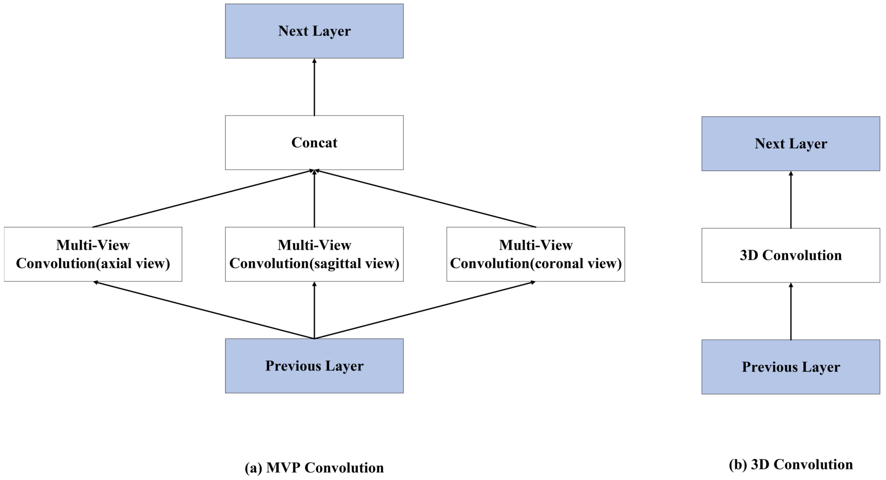
The images are preprocessed according to the following three steps before fed into the proposed MVP U-Net. First, each image is cropped to the region of nonzero values, and, at the same time, the image is normalized to the [2.0, 98.0] percentiles of the intensity values of the entire image. Second, the brain regions of images for each modality are normalized by Z-score normalization. The region outsides the brain is set to 0. Third, batch generators (a python package maintained by the Division of Medical Image Computing at the German Cancer Research Center) are applied to do data augmentation, including random elastic deformation, rotation, scaling, and mirroring [12].

2.2 Network Architecture

MVP Convolution Block. The architecture of the proposed MVP convolution is shown in Fig. 1(a), where we divide a 3D convolution into three orthogonal views (axial, sagittal, coronal) in a parallel fashion, followed by a pointwise convolution. The pointwise convolution is part of the Depthwise Separable Convolution network first proposed by Google [7], which consists of a depthwise convolution and a pointwise convolution. Figure 1(b) shows the traditional 3D convolution.

Figure 2 shows our MVP convolution block, which includes an activation function and an instance normalization, where our activation function is LeakyReLU (leakiness = 0.01). At the same time, in order to solve the problem of gradient disappearance caused by the increase of depth, we add the residual structure on the basis of the original structure. MVP convolution block is the main contribution of our proposed method. Each level of the network comprises two MVP convolution blocks of different resolutions.

MVP U-Net Architecture. The proposed MVP U-Net follows the encoder-decoder based 3D U-Net architecture. Instead of traditional 3D convolution [19], we employ the multi-view convolution to learn spatial-temporal features and the pointwise convolution to learn channel features. The multi-view convolution performs 2D convolutions in three orthogonal views of the input data, i.e., axial, sagittal, coronal. The pointwise convolution [10] is used to merge the antecedent outputs. In this way, the generalization ability of the model can be improved while the number of parameters can be reduced.


Next Layer
Concat
Multi-View Convolution(axial view)
Multi-View Convolution(sagittal view)
Multi-View Convolution(coronal view)
Previous Layer
Next Layer
3D Convolution
Previous Layer

(a) MVP Convolution

(b) 3D Convolution

Fig. 1. Comparison of MVP convolution and 3D convolution.

The sketch map of the proposed network is shown in Fig. 3. Just like the original 3D U-Net [8], our network consists of three parts: 1) the left part corresponds to the contracting path that encodes the increasingly abstract representation of the input. Different layers are connected through an encoder module which consists of a $3 \times 3 \times 3$ convolution with stride 2, padding 1 instead of max pooling; 2) the right part corresponds to the expanding path that restores the original resolution, and 3) the jump connection which corresponds to connecting the encoder results to the output of submodules with the same resolution in the encoder as input to the next submodule in the decoder.

MVP U-Net with the SE Block Architecture. SE block consists of three operation modules: Squeeze, Exception, and Reweight. It is a new subunit structure which focuses on the characteristic channel. Among them, the Squeeze operation is to compress each feature channel in the spatial dimension and transform each two-dimensional feature channel into a real number. The real number has the global receptive field to some extent, and the output dimension matches the number of input feature channels. The Exception operation is a mechanism similar to the gate in recurrent neural networks. It can generate weights for each feature channel, which is learned to explicitly model the correlation between feature channels. The Reweight operation regards the output weight of the exception operation as the importance of each feature channel and then weights it to the previous feature channel by channel through multiplication. After the above steps, the recalibration of the original features in the channel dimension is completed.

However, it was originally proposed to improve the classification performance of two-dimensional images. We modify it properly so that it can be used in the

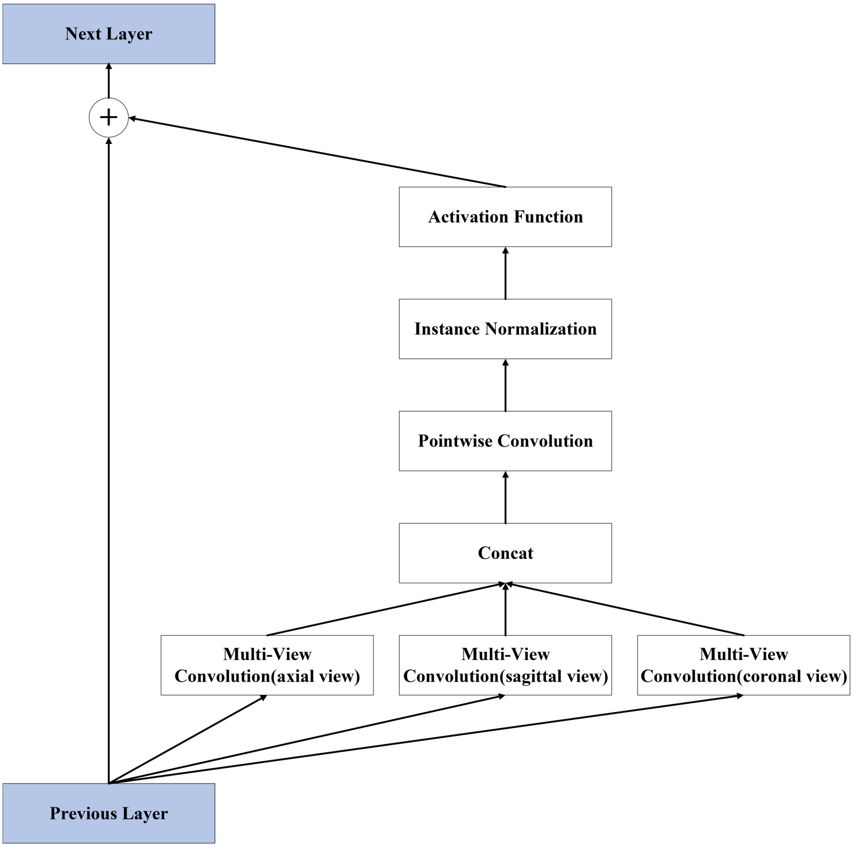

Next Layer
+
Activation Function
Instance Normalization
Pointwise Convolution
Concat
Multi-View Convolution(axial view)
Multi-View Convolution(sagittal view)
Multi-View Convolution(coronal view)
Previous Layer

Fig. 2. The architecture of the proposed MVP convolution block.

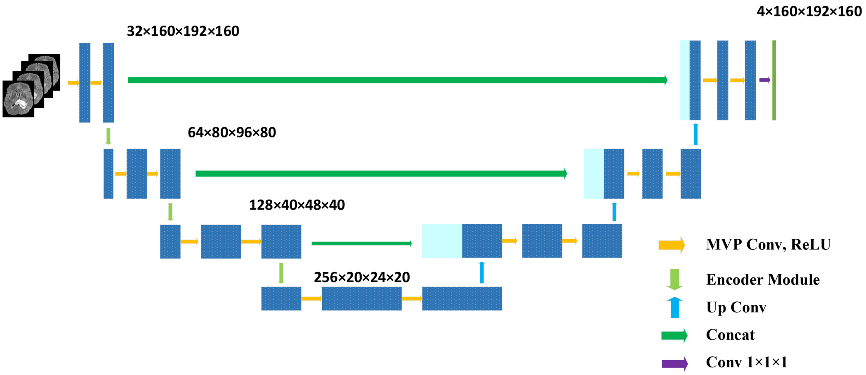


Fig. 3. The architecture of the proposed MVP U-Net.

classification of 3D feature map, and introduce it into our MVP U-Net after the concatenation section, as is shown in Fig. 4. It gives different weights to the features of different channels in the feature map after concatenation, in order to enhance those related features and suppress those less related features.

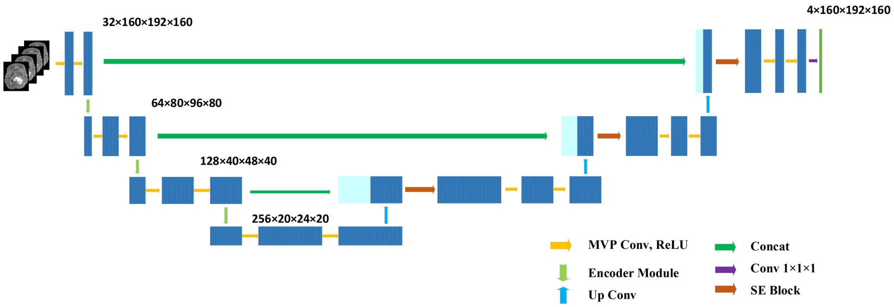


Fig. 4. The architecture of the proposed MVP U-Net with SE block.

2.3 Loss

The performance of a neural network depends not only on the choice of the network structure but also on the choice of the loss function, especially in the case of class imbalance. It holds for the task of brain tumor segmentation, in which the dataset varies in the size of classes [5,14]. In this paper, a hybrid loss function is employed that combines a multiclass Dice loss, used for multi-classification segmentation, and a focal loss aimed to alleviate class imbalance. Our loss function can be expressed as follows,

$$L = L_{Dice} + L_{focal} \tag{1}$$

The Dice loss is defined as,

$$L_{Dice} = \left(1 - \frac{2}{K}\sum_{k \in K} \frac{\sum_i u_i^k v_i^k}{\sum_i u_i^k + \sum_i v_i^k}\right) \tag{2}$$

where u is the softmax of the output map, v is the one-hot encoding of the corresponding ground truth label, i is the number of voxels of the output map and the corresponding ground truth label, k represents the current class, and K is the total number of classes.

The focal loss [16] is defined as,

$$L_{focal} = \begin{cases} -\alpha(1-y')^{\gamma} \log y' & , \quad y = 1 \\ -(1-\alpha)y'^{\gamma} \log(1-y') & , \quad y = 0 \end{cases} \tag{3}$$

where α and γ are constants. In our experiments, they are 0.25 and 2, respectively. y is the voxel value of the output map, and correspondingly, y' is the voxel value of the ground truth label.

2.4 Optimization

We use Adam optimizer to train our model [13]. The learning rate decreases as the epoch increases, which can be expressed as

$$lr = lr_0 * (1 - \frac{i}{N_i})^{0.9} \tag{4}$$

where i represents the current number of epochs, N_i is total number of epochs. The initial learning rate lr_0 is set as 10^{-4}.

3 Experiments and Results

We use the data provided by Brain Tumor Segmentation (BraTS) Challenge 2020 to evaluate the proposed network. The training dataset consists of 369 cases with accompanying ground truth labels by expert board-certified neuroradiologists. Our model is trained on one GeForce GTX 1080Ti GPU in a Pytorch environment. The batch size is 1 and the patch size is set to 160 × 192 × 160. We concatenate four modalities into a four-channel feature map as input where each channel represents one modality. The results of our MVP U-Net on the BraTS 2020 training dataset are shown in Table 1, and the BraTS 2020 Training 013 case in training dataset with groundtruth and predicted labels are shown in Fig. 5.

Table 1. Mean Dice, Hausdorff95, Sensitivity and Specificity on BraTS 2020 training dataset of the proposed method: original MVP U-Net. ET: enhancing tumor, WT: whole tumor, TC: tumor core.

	Dice			Sensitivity			Specificity			Hausdorff95		
	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
Original MVP U-Net	0.600	0.799	0.635	0.676	0.909	0.716	0.999	0.997	0.999	56.655	29.831	26.878

The validation dataset and testing dataset contain 125 and 166 cases with unknown glioma grade and unknown segmentation, respectively. Ground truth segmentations for them are unknown and the evaluation is carried out via an online CBICA portal for the BraTS 2020 challenge. The models we have trained on the training dataset, including the original 3D U-Net, the original MVP U-Net, and MVP U-Net with SE block, are respectively used to predict the validation dataset of BraTS 2020, and the quantitative evaluation is obtained as shown in Table 2. As can be seen from the results, compared with the original 3D U-Net, the original MVP U-Net and the MVP U-Net with SE block has improved performance in most metrics. Meanwhile, the segmentation effect of the MVP U-Net with SE block is better than the original MVP U-Net.

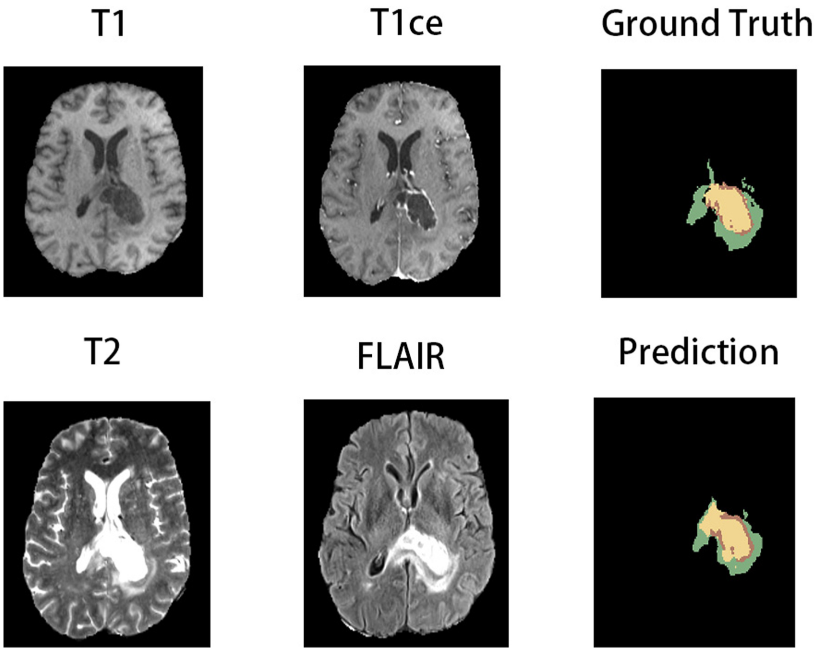


Fig. 5. The BraTS 2020 Training 013 case of training dataset with groundtruth and predicted labels (yelow:NCR/NET, green:ED, red:ET). (Color figure online)

Table 2. Mean Dice, Hausdorff95, Sensitivity and Specificity on BraTS 2020 validation dataset of the proposed methods: original MVP U-Net and MVP U-Net with SE block. ET: enhancing tumor, WT: whole tumor, TC: tumor core.

	Dice			Sensitivity			Specificity			Hausdorff95		
	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
Original 3D U-Net	0.585	0.762	0.604	0.686	0.868	0.742	0.999	0.997	0.998	78.429	43.598	44.543
Original MVP U-Net	0.601	0.785	0.639	0.659	0.901	0.715	0.993	0.970	0.997	56.653	29.837	26.870
MVP U-Net with the SE block	0.671	0.862	0.623	0.675	0.885	0.634	1.000	0.998	0.999	47.333	12.581	50.149

Finally, we used the MVP U-Net with the SE block to predict the testing dataset, and the results are shown in Table 3. Our method achieves average Dice scores of 0.715, 0.839, and 0.768 for enhancing tumor, whole tumor, and tumor core, respectively. The results are similar to those in the validation dataset, indicating that the model we designed has achieved desirable results in the automatic segmentation of multimodal brain tumors and the generalization ability of this model is also relatively powerful.

Table 3. Dice, Hausdorff95, Sensitivity and Specificity on BraTS 2020 testing dataset of the proposed method: MVP U-Net with SE block. ET: enhancing tumor, WT: whole tumor, TC: tumor core.

	Dice			Sensitivity			Specificity			Hausdorff95		
	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
Mean	0.715	0.839	0.768	0.756	0.912	0.800	0.999	0.998	0.999	33.147	10.362	33.577
StdDev	0.272	0.160	0.292	0.301	0.141	0.318	0.002	0.003	0.003	96.501	16.592	92.627
Median	0.818	0.893	0.892	0.903	0.961	0.962	1.000	0.998	1.000	2.000	4.243	3.317
25quantile	0.654	0.835	0.742	0.665	0.900	0.816	0.999	0.997	0.999	1.414	2.871	2.000
75quantile	0.895	0.927	0.940	0.957	0.985	0.988	1.000	0.999	1.000	4.472	7.729	9.312

4 Conclusion

In this paper, we propose a novel CNN-based neural network called Multi-View Pointwise (MVP) U-Net for brain tumor segmentation from multi-model 3D MRI. We use three multi-view convolutions and one pointwise convolution to reconstruct the 3D convolution in conventional 3D U-Net, in which the purpose of multi-view convolution is to learn spatial-temporal features while pointwise convolution to learn channel features. In this way, the proposed architecture can not only improve the generalization ability of the network but also reduce the number of parameters. Further, we modify the SE block properly and introduce it into our original MVP U-Net after the concatenation section. Experiments showed that the performance of this method was improved compared with the original MVP U-Net.

During the experiment, we tried a variety of approaches. We found that the model performance could be improved by changing the encoders of the U-shaped network from max pooling to 3D convolution, and the results could also be improved by increasing the number of channels. Finally, the trained MVP U-Net with SE block was used to predict the testing dataset, and achieved mean Dice scores of 0.715, 0.839, and 0.768 for enhancing tumor, whole tumor, and tumor core, respectively. The results showed the effectiveness of the proposed MVP U-Net with the SE block for multi-modal brain tumor segmentation.

In the future, we will make further efforts in data preprocess and network architecture design to alleviate the imbalance of tumor categories and improve the accuracy of tumor segmentation.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant No. 61903336, 61703369, 61976190, Natural Science Foundation of Zhejiang Province under Grant No. LY21F030015, Key Research and Development Program of Zhejiang Province under Grant No. 2020C03070, Major Science & Technology Projects of Wenzhou under Grant No ZS2017007.

References

1. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The cancer imaging archive. Nat. Sci. Data **4**, 170117 (2017)
2. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Cancer Imaging Arch. **286** (2017)
3. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
4. 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)
5. Berman, M., Triki, A.R., Blaschko. , M.BT:he lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4413–4421 (2018)
6. Chen, W., Liu, B., Peng, S., Sun, J., Qiao, X.: S3D-UNet: separable 3D U-Net for brain tumor segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 358–368. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_32
7. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
8. 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
9. Haque, H., Hashimoto, M., Uetake, N., Jinzaki, M.: Semantic segmentation of thigh muscle using 2.5 d deep learning network trained with limited datasets. arXiv preprint arXiv:1911.09249 (2019)
10. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)d
12. Huang, C., Han, H., Yao, Q., Zhu, S., Zhou, S.K.: 3D U^2-Net: a 3D universal u-net for multi-domain medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 291–299. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_33
13. 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
14. Kamnitsas, K., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 450–462. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_38
15. Li, C., Zhong, Q., Xie, D., Pu, S.: Collaborative spatiotemporal feature learning for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7872–7881 (2019)

16. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
17. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 305–321 (2018)