QCResUNet: Joint Subject-Level and Voxel-Level Prediction of Segmentation Quality

Peijie Qiu[1](✉), Satrajit Chakrabarty[2], Phuc Nguyen[3], Soumyendu Sekhar Ghosh[2], and Aristeidis Sotiras[1,4]

[1] Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO, USA
peijie.qiu@wustl.edu
[2] Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO, USA
[3] Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH, USA
[4] Institute for Informatics, Data Science and Biostatistics, Washington University School of Medicine, St. Louis, MO, USA

Abstract. Deep learning has achieved state-of-the-art performance in automated brain tumor segmentation from magnetic resonance imaging (MRI) scans. However, the unexpected occurrence of poor-quality outliers, especially in out-of-distribution samples, hinders their translation into patient-centered clinical practice. Therefore, it is important to develop automated tools for large-scale segmentation quality control (QC). However, most existing QC methods targeted cardiac MRI segmentation which involves a single modality and a single tissue type. Importantly, these methods only provide a subject-level segmentation-quality prediction, which cannot inform clinicians where the segmentation needs to be refined. To address this gap, we proposed a novel network architecture called QCResUNet that simultaneously produces segmentation-quality measures as well as voxel-level segmentation error maps for brain tumor segmentation QC. To train the proposed model, we created a wide variety of segmentation-quality results by using i) models that have been trained for a varying number of epochs with different modalities; and ii) a newly devised segmentation-generation method called SegGen. The proposed method was validated on a large public brain tumor dataset with segmentations generated by different methods, achieving high performance on the prediction of segmentation-quality metric as well as voxel-wise localization of segmentation errors. The implementation will be publicly available at https://github.com/peijie-chiu/QC-ResUNet.

Keywords: Automatic quality control · Brain tumor segmentation · Deep Learning

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 173–182, 2023.
https://doi.org/10.1007/978-3-031-43901-8_17

1 Introduction

Gliomas are the most commonly seen central nervous system malignancies with aggressive growth and low survival rates [19]. Accurate multi-class segmentation of gliomas in multimodal magnetic resonance imaging (MRI) plays an indispensable role in quantitative analysis, treatment planning, and monitoring of progression and treatment. Although deep learning-based methods have achieved state-of-the-art performance in automated brain tumor segmentation [6–8,14], their performance often drops when tasked with segmenting out-of-distribution samples and poor-quality artifactual images. However, segmentations of desired quality are required to reliably drive treatment decisions and facilitate clinical management of gliomas. Therefore, tools for automated quality control (QC) are essential for the clinical translation of automated segmentation methods. Such tools can enable a streamlined clinical workflow by identifying catastrophic segmentation failures, informing clinical experts where the segmentations need to be refined, and providing a quantitative measure of quality that can be taken into account in downstream analyses.

Most previous studies of segmentation QC only provide subject-level quality assessment by either directly predicting segmentation-quality metrics or their surrogates. Specifically, Wang et al. [18] leveraged a variational autoencoder to learn the latent representation of good-quality image-segmentation pairs in the context of cardiac MRI segmentation. During the inference, an iterative search scheme was performed in the latent space to find a surrogate segmentation. This segmentation is assumed to be a good proxy of the (unknown) ground-truth segmentation of the query image, and can thus be compared to the at-hand predicted segmentation to estimate its quality. Another approach that takes advantage of the pairs of images and ground-truth segmentation is the reverse classification accuracy (RCA) framework [13,17]. In this framework, the test image is registered to a preselected reference dataset with known ground-truth segmentation. The quality of a query segmentation is assessed by warping the query image to the reference dataset. However, these methods primarily targeted QC of cardiac MRI segmentation, which involves a single imaging modality and a single tissue type with a welch-characterized location and appearance. In contrast, brain tumor segmentation involves the delineation of heterogeneous tumor regions, which are manifested through intensity changes relative to the surrounding healthy tissue across multiple modalities. Importantly, there is significant variability in brain tumor appearances, including multifocal masses and complex shapes with heterogeneous textures. Consequently, adapting approaches for automated QC of cardiac segmentation to brain tumor segmentation is challenging. Additionally, iterative search or registration during inference makes the existing methods computationally expensive and time-consuming, which limits their applicability in large-scale segmentation QC.

Multiple studies have also explored regression-based methods to directly predict segmentation-quality metrics, e.g., Dice Similarity Coefficient (DSC). For example, Kohlberger et al. [10] used Support Vector Machine (SVM) with handcrafted features to detect cardiac MRI segmentation failures. Robinson et al. [12]

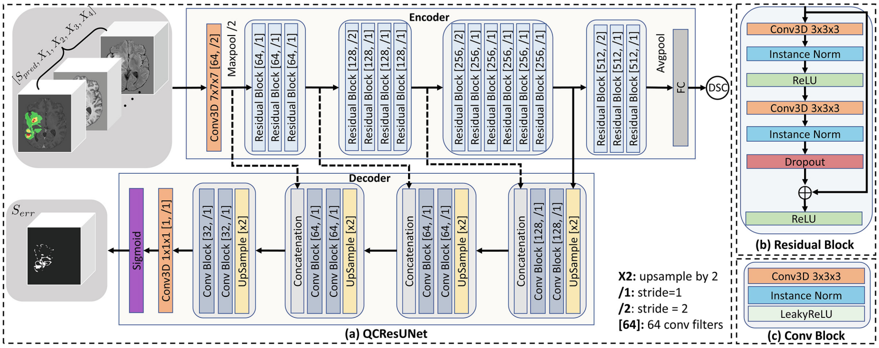

Fig. 1. (a) The proposed QCResUNet model adopts an encoder-decoder neural network architecture that takes four modalities and the segmentation to be evaluated. Given this input, QCResUNet predicts the DSC and segmentation error map. (b) The residual block in the encoder. (c) The convolutional block in the decoder.

proposed a convolutional neural network (CNN) to automatically extract features from segmentations generated by a series of Random Forest segmenters to predict DSC for cardiac MRI segmentation. Kofler et al. [9] proposed a CNN to predict holistic ratings of segmentations, which were annotated by neuroradiologists, with the goal of better emulating how human experts. Though these regression-based methods are advantageous for fast inference, they do not provide voxel-level localization of segmentation failures, which can be crucial for both auditing purposes and guiding manual refinements.

In summary, while numerous efforts have been devoted to segmentation QC, most works were in the context of cardiac MRI segmentation with few works tackling segmentation QC of brain tumors, which have more complex and heterogeneous appearances than the heart. Furthermore, most of the existing methods do not localize segmentation errors, which is meaningful for both auditing purposes and guiding manual refinement. To address these challenges, we propose a novel framework for joint subject-level and voxel-level prediction of segmentation quality from multimodal MRI. The contribution of this work is four-fold. First, we proposed a predictive model (QCResUNet) that simultaneously predicts DSC and localizes segmentation errors at the voxel level. Second, we devised a data-generation approach, called SegGen, that generates a wide range of segmentations of varying quality, ensuring unbiased model training and testing. Third, our end-to-end predictive model yields fast inference. Fourth, the proposed method achieved a good performance in predicting subject-level segmentation quality and identifying voxel-level segmentation failures.

2 Method

Given four imaging modalities denoted as $[X_1, X_2, X_3, X_4]$ and a predicted multiclass brain tumor segmentation mask (S_{pred}), the goal of our approach is to automatically assess the tumor segmentation quality by simultaneously predicting DSC and identifying segmentation errors as a binary mask (S_{err}). Toward this end, we proposed a 3D encoder-decoder architecture termed QCResUNet (see Fig. 1(a)) for simultaneously predicting DSC and localizing segmentation errors. QCResUNet has two parts trained in an end-to-end fashion: i) a ResNet-34 [4] encoder for DSC prediction; and ii) a decoder architecture for segmentation error map prediction (i.e., the difference between predicted segmentation and ground-truth segmentation).

The ResNet-34 encoder enables the extraction of semantically rich features that are useful for characterizing the quality of the segmentation. We maintained the main structure of the vanilla 2D ResNet-34 [4] but made the following modifications, which were necessary to account for the 3D nature of the input data (see Fig. 1(b)). First, all the 2D convolutional layers and pooling layers in the vanilla ResNet were changed to 3D. Second, the batch normalization [5] was replaced by instance normalization [16] to accommodate the small batch size in 3D model training. Third, spatial dropout [15] with a probability of 0.3 was added to each residual block to prevent overfitting.

The building block of the decoder consisted of an upsampling by a factor of two, which was implemented by a nearest neighbor interpolation in the feature map, followed by two convolutional blocks that halve the number of feature maps. Each convolutional block comprised a 3 × 3 × 3 convolutional layer followed by an instance normalization layer and a leaky ReLU activation [11] (see Fig. 1(c)). The output of each decoder block was concatenated with features from the corresponding encoder level to facilitate information flow from the encoder to the decoder. Compared to the encoder, we used a shallower decoder with fewer parameters to prevent overfitting and reduce computational complexity.

The objective function for training QCResUNet consists of two parts. The first part corresponds to the DSC regression task. It consists of a mean absolute error (MAE) loss ($\mathcal{L}_{MAE}$) term that penalizes differences between ground truth (DSC_{gt}) and predicted DSC (DSC_{pred}):

$$\mathcal{L}_{MAE} = \frac{1}{N}\sum_{n=1}^{N} |DSC_{gt}^{(n)} - DSC_{pred}^{(n)}|_1, \tag{1}$$

where N denotes the number of samples in a batch. The second part of the objective function corresponds to the segmentation error prediction. It consists of a dice loss [3] and a binary cross-entropy loss, given by:

$$\begin{aligned} \mathcal{L}_{dice} &= -\frac{2 \cdot \sum_{i=1}^{I} S_{err_{gt}}^{(i)} \cdot S_{err_{pred}}^{(i)}}{\sum_{i=1}^{I} S_{err_{gt}}^{(i)} + \sum_{i=1}^{I} S_{err_{pred}}^{(i)}} \\ \mathcal{L}_{BCE} &= -\frac{1}{I}\sum_{i=1}^{I} S_{err_{gt}}^{(i)} \log S_{err_{pred}}^{(i)} + (1 - S_{err_{gt}}^{(i)}) \log(1 - S_{err_{pred}}^{(i)}), \end{aligned} \tag{2}$$

where $S_{err_{gt}}$, $S_{err_{pred}}$ denote the binary ground-truth segmentation error map and the predicted error segmentation map from the sigmoid output of the

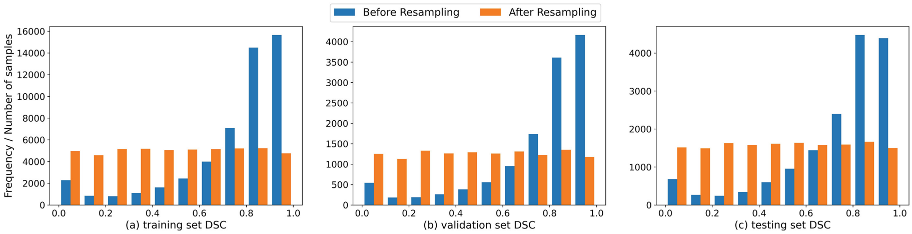

Fig. 2. DSC distribution of the generated dataset before and after resampling. (a), (b) and (c) are the DSC distribution for the training, validation, and testing set.

decoder, respectively. The dice loss and cross-entropy loss were averaged across the number of pixels I in a batch. The two parts are combined using a weight parameter λ to balance the different loss components:

$$\mathcal{L}_{total} = \mathcal{L}_{MAE} + \lambda\left(\mathcal{L}_{dice} + \mathcal{L}_{BCE}\right). \tag{3}$$

3 Experiments

For this study, pre-operative multimodal MRI scans of varying grades of glioma were obtained from the 2021 Brain Tumor Segmentation (BraTS) challenge [1] training dataset (n = 1251). For each subject, four modalities viz. pre-contrast T1-weighted (T1), T2-weighted (T2), post-contrast T1-weighted (T1c), and Fluid attenuated inversion recovery (FLAIR) are included in the dataset. It also included expert-annotated multi-class tumor segmentation masks comprising enhancing tumor (ET), necrotic tumor core (NCR), and edema (ED) classes. All data were already registered to a standard anatomical atlas and skull-stripped. The skull-stripped scans were then z-scored to zero mean and unit variance. All the data was first cropped to non-zero value regions, and then zero-padded to a size of 160 × 192 × 160 to be fed into the network.

3.1 Data Generation

The initial dataset was expanded by producing segmentation results at different levels of quality to provide an unbiased estimation of segmentation quality. To this end, we adopted a three-step approach. First, a nnUNet framework [6] was adopted and trained five times separately using different modalities as input (i.e., T1-only, T1c-only, T2-only, FLAIR-only, and all four modalities). As only certain tissue-types are captured in each modality (e.g., enhancing tumor is captured well in T1c but not in FLAIR), this allowed us to generate segmentations of a wide range of qualities. nnUNet was selected for this purpose due to its wide success in brain tumor segmentation tasks. Second, to further enrich our dataset with segmentations of diverse quality, we sampled segmentations along the training routines at different iterations. A small learning rate (1×10^{-6}) was chosen in training all the models to slower their convergence in order to sample segmentations gradually sweeping from poor quality to high quality. Third, we devised a method called SegGen that applied image transformations, including random rotation (angle $= [-15°, 15°]$), random scaling (scale $= [0.85, 1.25]$), random translation (moves $= [-20, 20]$), and random elastic deformation (displacement $= [0, 20]$), to the ground-truth segmentations with a probability of 0.5, resulting in three segmentations for each subject.

The original BraTS 2021 training dataset was split into training ($n = 800$), validation ($n = 200$), and testing ($n = 251$) sets. After applying the three-step approach, it resulted in 48000, 12000, and 15060 samples for the three sets, respectively. However, this generated dataset suffered from imbalance (Fig. 2(a), (b), and (c)) because the CNN models could segment most of the cases correctly. Training using such an imbalanced dataset is prone to producing biased models that do not generalize well. To mitigate this issue, we proposed a resampling strategy during the training to make the DSC more uniformly distributed. Specifically, we used the Quantile transform to map the distribution of a variable to a target distribution by randomly smoothing out the samples unrelated to the target distribution. Using the Quantile transform, the data generator first transformed the distribution of the generated DSC to a uniform distribution. Next, the generated samples closest to the transformed uniform distribution in terms of Euclidean distance were chosen to form the resampled dataset. After applying our proposed resampling strategy, the DSC in the training and validation set approached a uniform distribution (Fig. 2(a), (b), and (c)). The total number of samples before and after resampling remained the same with repeating samples. We kept the resampling stochastic at each iteration during training to make all the generated samples seen by the model. The generated testing set was also resampled to perform an unbiased estimation of the quality at different levels resulting in 4895 samples.

In addition to the segmentations generated by the nnUNet framework and the SegGen method, we also generated out-of-distribution segmentation samples for the testing set to validate the generalizability of our proposed model. For this purpose, five models were trained on the training set using the DeepMedic framework [8] with different input modalities (i.e., T1-only, T1c-only, T2-only, FLAIR-only, and all four modalities). This resulted in $251 \times 5 = 1255$ out-of-distribution samples in the testing set.

3.2 Experimental Design

Baseline Methods: In this study, we compared the performance of the proposed model to three baseline models: (i) a UNet model [14], (ii) a ResNet-34 [4], and (iii) the ReNet-50 model used by Robinson et al. [12]. For a fair comparison, the residual blocks in the ResNet-34 and ResNet-50 were the same as that in the QCResUNet. We added an average pooling followed by a fully-connected layer to the last feature map of the UNet to predict a single DSC value. The evaluation was conducted on in-sample (nnUNet and SegGen) and out-of-sample segmentations generated by DeepMedic.

Table 1. The QC performance of three baseline methods and the proposed method was evaluated on in-sample (nnUNet and SegGen) and out-of-sample (DeepMedic) segmentations. The best metrics in each column are highlighted in bold. DSC$_{err}$ denotes the median DSC between $S_{err_{gt}}$ and $S_{err_{pred}}$ across all samples.

Method	In-sample (nnUNet+SegGen)			Out-of-sample (DeepMedic)		
	r	MAE	DSC$_{err}$	r	MAE	DSC$_{err}$
UNet	0.895	0.1106 ± 0.085	–	0.923	0.0835 ± 0.063	–
ResNet34	0.942	0.0711 ± 0.063	–	0.942	0.0751 ± 0.065	–
ResNet50 [12]	0.944	0.0662 ± 0.065	–	0.943	0.0782 ± 0.067	–
QCResUNet(Ours)	**0.964**	$\mathbf{0.0570 \pm 0.050}$	**0.834**	**0.966**	$\mathbf{0.0606 \pm 0.049}$	**0.867**

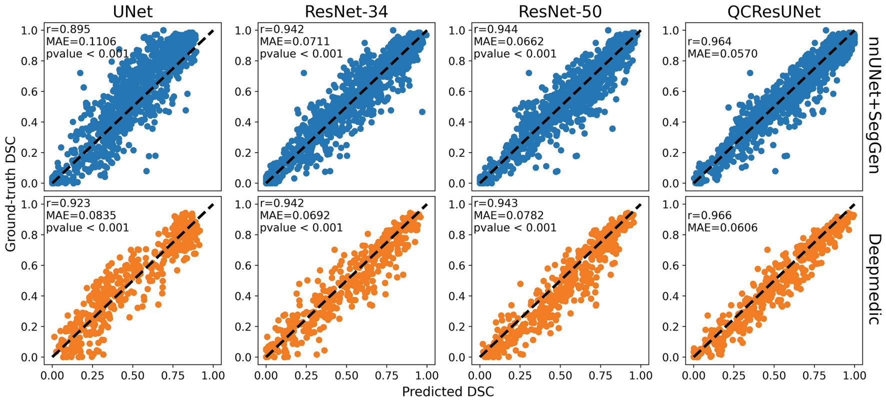

Fig. 3. Comparison of QC performance between three baseline methods (UNet, ResNet-34, ResNet-50) and the proposed method (QCResUNet) for segmentations generated using nnUNet and SegGen (top row) as well as DeepMedic (bottom row).

Training Procedure: All models were trained for 150 epochs using an Adam optimizer with a L_2 weight decay of 5×10^{-4}. The batch size was set to 4. Data augmentation, including random rotation, random scaling, random mirroring, random Gaussian noise, and Gamma intensity correction, was applied to prevent overfitting during training. We performed a random search [2] to determine the optimal hyperparameters (i.e., initial learning rate and loss weight balance parameter λ) on the training and validation set. The hyperparameters that yielded the best results were $\lambda = 1$ and an initial learning rate of 1×10^{-4}. The learning rate was exponentially decayed by a factor of 0.9 at each epoch until 1×10^{-6}. Model training was performed on four NVIDIA Tesla A100 and V100S GPUs. The proposed method was implemented in PyTorch v1.12.1.

Statistical Analysis: We assessed the performance of the subject-level segmentation quality prediction in terms of Pearson coefficient r and MAE between the predicted DSC and the ground-truth DSC. The performance of the segmentation error localization was assessed by the DSC$_{err}$ between the predicted segmenta-

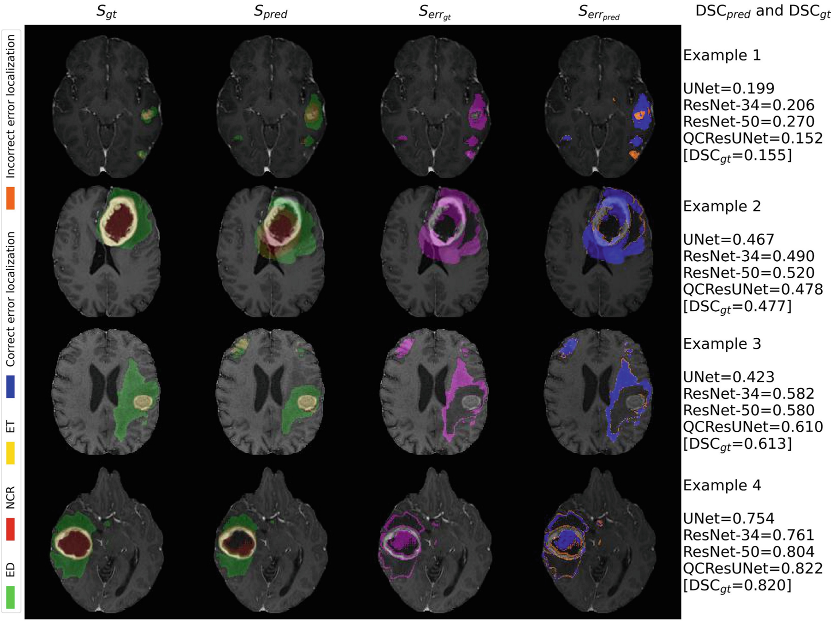

Fig. 4. Examples showcasing the performance of the proposed methods. The last column denotes the QC performance of different methods. The penultimate column denotes the predicted segmentation error.

tion error map and the ground-truth segmentation error map. P-values were computed using a paired t-test between DSC predicted by QCResUNet versus ones predicted by corresponding baselines.

4 Results

The proposed QCResUNet achieved good performance in predicting subject-level segmentation quality for in-sample (MAE = 0.0570, r = 0.964) and out-of-sample (MAE = 0.0606, r = 0.966) segmentations. The proposed method also showed statistically significant improvement against all three baselines (Table 1 and Fig. 3). We found that the DSC prediction error (MAE) of the proposed method was distributed more evenly across different levels of quality than all baselines (see Fig. 3) with a smaller standard deviation of 0.050 for in-sample segmentations and 0.049 for out-of-sample segmentations. A possible explanation is that the joint training of predicting subject-level and voxel-level quality enabled the QCResUNet to learn deep features that better characterize the segmentation quality. For the voxel-level segmentation error localization task, the model achieved a median DSC of 0.834 for in-sample segmentations and 0.867 for out-of-sample segmentations. This error localization is not provided by any of the baselines and enables QCResUnet to track segmentation failures at different levels of segmentation quality (Fig. 4).

5 Conclusion

In this work, we proposed a novel CNN architecture called QCResUNet to perform automatic brain tumor segmentation QC in multimodal MRI scans. QCResUNet simultaneously provides subject-level segmentation-quality prediction and localizes segmentation failures at the voxel level. It achieved superior DSC prediction performance compared to all baselines. In addition, the ability to localize segmentation errors has the potential to guide the refinement of predicted segmentations in a clinical setting. This can significantly expedite clinical workflows, thus improving the overall clinical management of gliomas.

Acknowledgements. All computations were supported by the Washington University Center for High Performance Computing, which was partially funded by NIH grants S10OD025200, 1S10RR022984-01A1, and 1S10OD018091-01.

References

1. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**(2) (2012)
3. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: Carneiro, G., et al. (eds.) LABELS/DLMIA -2016. LNCS, vol. 10008, pp. 179–187. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46976-8_19
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
7. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 287–297. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_25
8. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)
9. Kofler, F., et al.: Deep quality estimation: creating surrogate models for human quality ratings. arXiv preprint arXiv:2205.10355 (2022)
10. Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L.: Evaluating segmentation error without ground truth. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7510, pp. 528–536. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33415-3_65

11. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the ICML, Atlanta, Georgia, USA, vol. 30, p. 3 (2013)
12. Robinson, R., et al.: Real-time prediction of segmentation quality. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 578–585. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_66
13. Robinson, R., et al.: Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. J. Cardiovasc. Magn. Reson. **21**(1), 1–14 (2019)
14. 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
15. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656 (2015)
16. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
17. Valindria, V.V., et al.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Trans. Med. Imaging **36**(8), 1597–1606 (2017)
18. Wang, S., et al.: Deep generative model-based quality control for cardiac MRI segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 88–97. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_9
19. Zhuge, Y., et al.: Brain tumor segmentation using holistically nested neural networks in MRI images. Med. Phys. **44**(10), 5234–5243 (2017)